UNIVERSITÉ DE MONTRÉAL

CONCEPTION D'UN SYSTÈME D'INFORMATION POUR SOUTENIR L'ANALYSE DES IMPACTS D'UNE NOUVELLE INFRASTRUCTURE DE TRANSPORT

JULIEN FAUCHER
DÉPARTEMENT DES GÉNIES CIVIL, GÉOLOGIQUE ET DES MINES
ÉCOLE POLYTECHNIQUE DE MONTRÉAL

MÉMOIRE PRÉSENTÉ EN VUE DE L'OBTENTION
DU DIPLÔME DE MAÎTRISE ÈS SCIENCES APPLIQUÉES
(GÉNIE CIVIL)
DÉCEMBRE 2013

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Ce mémoire intitulé :

CONCEPTION D'UN SYSTÈME D'INFORMATION POUR SOUTENIR L'ANALYSE DES IMPACTS D'UNE NOUVELLE INFRASTRUCTURE DE TRANSPORT

présenté par : FAUCHER Julien
en vue de l'obtention du diplôme de : Maîtrise ès sciences appliquées
a été dûment accepté par le jury d'examen constitué de :

M. TRÉPANIER Martin, Ph.D., président
Mme MORENCY Catherine, Ph.D., membre et directrice de recherche
M. SAUNIER Nicolas, Ph.D., membre et codirecteur de recherche
M. PATTERSON Zachary, Ph.D., membre

REMERCIEMENTS

Je tiens à adresser mes remerciements les plus sincères à mes directeurs de recherche, les professeurs Catherine Morency et Nicolas Saunier de m'avoir permis de travailler sur un projet aussi complexe et enrichissant et pour m'avoir soutenu lors de sa réalisation. Leurs conseils, autant dans la direction générale du projet que sur des aspects plus techniques ont été d'un apport inestimable. Je leur suis aussi reconnaissant de m'avoir offert la chance de travailler, en marge de mes activités de recherche principales, sur d'autres projets qui se sont eux aussi révélés extrêmement enrichissants et formateurs.

J'aimerais aussi remercier l'ensemble des chercheurs liés au projet de suivi du nouveau pont de l'autoroute 25, qui ont apporté un soutien méthodologique dans les phases d'analyses. Finalement, merci à Philippe Gaudette, qui a participé à l'organisation de certains ensembles de données dans le cadre de travaux liés à son stage.

Merci à Pierre-Léo, qui m'a fourni quantité de soutien technique et des suggestions de solutions à d'innombrables problèmes. Finalement, je tiens à remercier les autres étudiants de transport, plus particulièrement Louiselle et Sébastien, qui ont été à la fois source de motivation et d'inspiration tout au long de ce projet.

RÉSUMÉ

La construction et la mise en service d'un nouveau pont dans le cadre du parachèvement du corridor de l'autoroute 25, entraînent une modification des comportements de mobilité de la population de la grande région de Montréal. Or, la quantification des effets de la mise en place d'une telle infrastructure de transport sur la circulation et sur l'environnement général est un processus rarement entrepris, si bien que les impacts réels restent méconnus. Le décret gouvernemental autorisant la construction du pont impose de procéder à de tels travaux d'observation de ses effets sur différents aspects, le tout dans une approche longitudinale afin de tracer l'évolution de ces impacts. L'évaluation précise des conséquences de la mise en service du pont repose toutefois sur des quantités très importantes de données colligées non seulement aux abords de la nouvelle infrastructure, mais aussi sur d'autres infrastructures majeures de transport et à de nombreux autres points d'intérêt du territoire.

La réalisation d'études nécessite donc un travail majeur d'organisation de l'information disponible afin d'offrir des assises solides sur lesquelles baser les analyses. L'objectif principal de ce projet est de créer un système d'information cohérent et flexible permettant de fusionner l'ensemble des données obtenues dans le cadre du mandat d'analyse de l'état de référence précédant la mise en service du nouveau pont. Les données présentant différents problèmes de qualité, celles-ci doivent dans un premier temps être documentées afin d'assurer la qualité et une standardisation de l'information. Cette standardisation est essentielle dans le but de réaliser une intégration de toutes les données obtenues. Suite à cette intégration, des travaux sommaires d'utilisation automatisée devraient permettre l'obtention rapide de résultats dans un format désiré qui permettra finalement de soutenir efficacement les analyses en plus d'assurer que la solution développée répond aux besoins.

Concrètement, les travaux énoncés dans ce mémoire se basent sur des ensembles de données fournis par différents organismes gouvernementaux, notamment le Ministère des Transports du Québec (MTQ), le Réseau de Surveillance de la Qualité de l'Air de la ville de Montréal (RSQA) et le Ministère du Développement Durable, de l'Environnement et des Parcs (MDDEP), pour une période de temps couvrant les années qui précèdent la mise en service du pont. Ces données comportent des mesures de qualité de l'air, de météo et de circulation dans un ensemble hétéroclite de formats présentant chacun des défis d'intégration propres qui doivent être documentés avant qu'une utilisation puisse en être faite. Certains des problèmes structurels et des défis d'intégration portent sur des niveaux de résolution spatio-temporels variables entre les ensembles de données et même parfois à l'intérieur d'un seul ensemble de données. Les codifications de différentes informations y sont aussi variables

et nécessitent une attention particulière.

L'organisation de l'information passe principalement par l'identification de structures communes à toutes les données rendues disponibles et par le développement d'un schéma permettant de structurer l'information dans un format commun, standard et normalisé. La définition du schéma se fait dans une approche orientée-objet se rapprochant de la réalité, le paradigme de collecte des données avec des stations physiques récoltant des mesures sur des paramètres précis étant relativement semblables à ce point pour toutes les données rendues disponibles. Ce point d'accès unique aux données doit par la suite permettre de rendre facile l'accès aux informations dans le but d'en faire l'analyse.

Afin de valider le modèle d'organisation de l'information développé, une implantation du schéma est faite dans un ensemble de solutions logicielles. Ces solutions doivent notamment permettre l'accès distant aux données et de stocker des informations géographiques afin de maximiser le potentiel du système d'information. Ces deux contraintes ont dirigé le choix vers l'implantation dans le système de base de données relationnelle PostgreSQL et ses extensions géographiques et le langage de programmation Ruby.

L'implantation nécessite la programmation d'outils effectuant l'insertion des données disponibles, en plus d'effectuer des manœuvres distinctes sur chaque enregistrement si nécessaire afin de résoudre les problèmes spécifiques identifiés pour chaque jeu de données. Par ailleurs, la mise en place d'un système de contraintes permet d'assurer la cohérence des informations en éliminant les risques de dédoublements des mesures ou des autres objets identifiés. Ces contraintes assurent par conséquent un contrôle de qualité de l'information non seulement dans le cadre des expériences réalisées au cours de ce mémoire mais devrait aussi garantir l'intégrité de l'information pour plusieurs phases d'insertion d'informations tout au long du programme de suivi des effets du nouveau pont.

Les travaux détaillés dans le cadre de ce mémoire visent donc principalement la mise en place du système d'information ainsi que d'en vérifier la pertinence et la capacité à soutenir les analyses pour la durée des mandats d'analyse des effets de la nouvelle infrastructure. Dans le but de s'assurer que le système peut remplir son objectif, une série d'outils visant à permettre l'analyse de l'état de référence, mais aussi à explorer les potentialités d'analyses sont développés et permettent de visualiser sous forme de tableaux, de cartes ou de graphiques les données obtenues, et ce, de façon automatisée. Les analyses faites à partir de ces tableaux et graphiques ont d'ailleurs contribué à faire l'évaluation de l'état de référence par une équipe de chercheurs de Polytechnique Montréal, de l'Université McGill et de l'Université Concordia. Finalement, la viabilité du système d'information à couvrir de plus longues périodes d'analyse est vérifiée grâce à des tests de performance, qui permettent de conclure que le système développé aura la capacité de produire les résultats attendus sans problèmes jusqu'à la fin

des mandats d'analyse, les performances restant stables pour des bases de données plus de cinq fois plus grandes que les données de la période de référence.

La réalisation de ce projet permet d'affirmer que le développement d'un système d'information peut fournir un soutien efficace aux analyses lorsque celles-ci se basent sur des ensembles de données de sources et de types variés. Néanmoins, la présence de nombreux cas spécifiques non traités en amont, ainsi que le problème de consolider des informations présentant des standards et des codifications aussi variées engendre des retard, si bien qu'une telle solution peut difficilement répondre aux objectifs si ceux-ci impliquent une production de résultats rapides.

ABSTRACT

The construction and opening of a new bridge in the larger scope of the completion of the highway 25 corridor has effects on travelers behavior in the greater Montreal region. The quantification of the effects engendered by a new infrastructure is a process that is rarely done and the global influence of such a structure on traffic and the environment is still calculated mostly through models and estimations. The government decree that authorized the building of the new bridge includes provisions to ensure that follow up studies be performed on a year by year basis to better understand the influence of the infrastructure. These post-opening evaluations rely on large quantities of data that are accumulated on a number of locations close to the corridor of the highway but also on other major roads as well of other points of interest.

The amount of data required to evaluate clearly the impacts of the new bridge leads to a need to organize and standardize the information clearly. The main objective of the project discussed in this memoir is to address this need for data organization through the development of a coherent and flexible information system that will be able to integrate all the data gathered during what is now considered the reference state preceding the opening of the bridge. However, some problems arise as many of the obtained datasets have consistency and quality issue that require investigation and documentation processes to assure data quality and eventually allow the definition of standards applicable to all the available data. The integration of all the information is then to be followed by some basic automation work that serve as a proof of concept by generating material supporting the analysis of the impacts.

Available data for the evaluation of the reference state contains air quality, weather and traffic readings from multiple stations operated by the Ministère des Transports du Québec (MTQ), the Réseau de Surveillance de la Qualité de l'Air de la ville de Montréal (RSQA) and the Ministère du Développement Durable, de l'Environnement et des Parcs (MDDEP). Some stations cover a short period before the opening of the bridge, while others are providing informations as far back as 2008. The formats in which the datasets are provided are extremely variable and heterogeneous, in form as well as in codification, which complicates the process of standardizing the information. Namely, a lot of the datasets provided present unique challenges that need to be resolved on a case by case basis. Problems identified range from basic time codification problems to resolution issue where stations present measures on varying intervals of time, often for the same data collection point.

The definition of an integrated information system relies on common structural patterns that have to be identified within the datasets provided. The information provided at this

point is determined to be driven by single general paradigm, although with slight variations, where a number of physical stations register measures of specific parameters over a defined period of time. Using an object-oriented approach, it is possible to define a data model that closely resembles the physical reality and that is able to provide storage points to all the available data. The integration of all the data in a single schema is beneficial in providing a single access point to all the obtained information, which presents the same structure and thus simplify the work of anyone who needs to access it.

The implantation of the model and the subsequent insertion of the data in a series of selected software were used to validate its capabilities to answer data storage needs. Although the design of the information system could be implemented in a number of acceptable software, the choice was made to use the Relational DataBase Management System (RDBMS) PostgreSQL and its geographical extensions as well as the Ruby programming language with the Ruby on Rails framework. This combination allows the storing of geographical objects and also provides capabilities for remote access. All of the provided data was inserted in the developed solution, including specific methods used to fix every dataset specific issues that were identified. A series of constraints were also developed to insure data quality and integrity and to avoid the creation of multiple entries in the hope of keeping the database as clean and organized as possible.

A set of tools were developed to integrate data in the system, as well as to provide basic output in pertinent form to support analysis. After this project, the developed tools are able to insert all of the provided data feeds in the information system, as well as provide tables, graphs and maps automatically that can support fact finding. The results of those generated visual representations were helpful in obtaining information to define the air quality and traffic situation prior to the opening of the bridge by a team of researchers from Polytechnique Montréal, McGill University and Concordia University. Finally, work was done to insure that the developed solution would be able to sustain the increasing size of the provided datasets and therefore support analysis over the next years of monitoring of the effects of the bridge. The conclusion of those analysis showed that the system can manage datasets five times larger than those that are available at this point, and therefore should still be effective over its planned lifespan.

This projects proves that a unified information system joining large and heterogeneous datasets can be useful in supporting analysis of a transportation infrastructure. However, the numerous problems identified within the datasets as well as the multiplication of their formats and standards both lead to lost time. Therefore, while it can definitely serve its purpose, such a system may not be able to support analysis if they need to be produced within a short time frame.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES EXTRAITS DE CODE

LISTE DES SIGLES ET ABRÉVIATIONS

POPE	Post-Opening Project Evaluation
MTQ	Ministère des Transports du Québec
MDDEP	Ministère du Développement Durable, de l'Environnement et des Parcs
SIG-T	Système d'Information Géographique appliqué aux transports
GIS-T	Geographical Information System for Transportation
RDBMS	Relational DataBase Management System
HAP	Hazardous Air Polluants
SGBD	Système de Gestion de Base de Données
SGBDR	Système de Gestion de Base de Données Relationnelle
GPS	Global Positioning System
RSQA	Réseau de Surveillance de la Qualité de l'Air de la ville de Montréal
COV	Composés Organiques Volatils
MOR	Mapping objet-relationnel
SIG	Système d'Information Géographique

LISTE DES ANNEXES

CHAPITRE 1

INTRODUCTION

La mise en place de nouvelles infrastructures routières se traduit inévitablement par des changements de comportements de la part des usagers, notamment en ce qui concerne les choix modaux, les itinéraires de déplacement ou encore les heures auxquelles ces déplacements sont faits. Les changements de comportements peuvent aussi entraîner des effets environnementaux en modifiant la localisation des émissions des polluants gazeux. Les variations des flots de circulation et de la fluidité du trafic ont alors le potentiel d'engendrer une variation du total de ces émissions. Or, l'analyse des impacts de nouvelles infrastructures et la comparaison avec les hypothèses de départ menant à leur réalisation restent des activités qui ne sont mises en place que très rarement.

Dans le cadre du projet de parachèvement de l'Autoroute 25, le gouvernement du Québec a autorisé, par le décret 1243-2005, la construction de tronçons autoroutiers ainsi que d'un nouveau pont entre l'île Jésus et l'île de Montréal, celui-ci a été construit dans le cadre d'un partenariat public-privé qui introduit des péages pour les utilisateurs de l'infrastructure. Selon les règles du décret, la mise en service du pont Olivier-Charbonneau, le 23 mai 2011, doit faire l'objet d'un suivi et d'une évaluation de ses impacts. Notamment, le décret demande l'organisation d'un programme de suivi de la qualité de l'air à proximité de l'axe de l'autoroute 25 dans un secteur allant de la jonction avec l'autoroute 440 au nord jusqu'au tunnel Louis-Hyppolite Lafontaine. La condition 8 du décret stipule plus spécifiquement que "Le programme de suivi, accompagné d'un état de référence, doit permettre de connaître la contribution du transport routier à la dégradation de la qualité de l'air ambiant".

C'est dans le cadre de la construction de cette infrastructure et dans l'intention d'en mesurer les effets qu'une équipe de chercheurs des Universités McGill, Concordia et de Polytechnique Montréal s'est vu confier la responsabilité de fournir des expertises scientifiques. Les mandats de cette équipe comprennent notamment l'identification des dépassements aux normes de qualité de l'air, l'étude des variations temporelles et spatiales des concentrations de divers polluants atmosphériques et une évaluation régionale des indices de qualité de l'air et de l'étendue de l'influence des nouvelles infrastructures. L'analyse de la contribution des transports aux dépassements ainsi que les corrélations entre les différents paramètres devront aussi être complétées afin d'identifier précisément les causes des variations. Les travaux pouvant répondre au mandat reposent sur des ensembles de données bruts de très grandes tailles aux origines diverses et présentant des problèmes de qualité et de cohérence importants. Dans

le but de pouvoir réaliser l'étude des effets du pont et de répondre aux mandats de recherche, il est nécessaire de construire un système d'information et d'effectuer des travaux de validation, de standardisation et de contrôle de qualité sur ces données afin de les organiser dans un tout cohérent.

1.1 Problématique de recherche

L'exploitation efficace d'ensembles de données aux structures variables provenant de sources multiples demande des travaux majeurs de planification et d'organisation de l'information. Dans le cas des informations entourant la mise en place de nouvelles infrastructures, la littérature étudiée survole généralement assez rapidement ces étapes d'organisation. Or, dans le cadre d'un projet de grande envergure visant à quantifier les effets longitudinaux d'une nouvelle infrastructure sur plusieurs années, l'accès à des données de qualité et organisées de façon claire et constante est un besoin essentiel.

La recherche porte donc sur l'organisation de ces données ainsi que sur l'élaboration d'un système d'information rassemblant des données de sources diverses permettant de produire des analyses des impacts de la mise en service de la nouvelle infrastructure sur le milieu environnant. Un tel système, d'abord basé sur des données de la période de référence, devrait être extensible afin de permettre d'y greffer des données d'exploitation ainsi que d'autres ensembles de données lorsque jugé pertinent et ainsi fournir une base solide afin de soutenir les analyses pour toute la durée du mandat.

1.2 Objectifs du projet de recherche

L'objectif principal du projet de recherche est la conception d'un système d'information cohérent et flexible qui permettra de fusionner les ensembles de données rendues disponibles dans le cadre des travaux d'évaluation de l'état de référence du nouveau pont. Un objectif secondaire est d'établir les bases d'analyse et de valider la qualité du système défini. Plusieurs objectifs spécifiques peuvent être énoncés :

— définir un système de gestion de données intégré et flexible ;

— joindre des ensembles de données de sources et de formats divers ;

— créer des outils de traitement automatisés ;

— définir des structures de contrôle assurant la qualité des données stockées ;

— démontrer le potentiel d'analyse de ces données.

1.3 Structure du document

Ce mémoire reprend l'ensemble des étapes requises dans le but de répondre à la liste des objectifs mentionnés précédemment. Le Chapitre 2 contient une revue de la littérature traitant des aspects théoriques associés à des variations de la capacité routière, des variations de comportements des utilisateurs du réseau routier, de plusieurs études de cas portant sur des changements de capacité routière ainsi que sur l'intégration des données. Le Chapitre 3 fait quant à lui état du contexte dans lequel s'insère le nouveau pont, définit les grandes lignes de la méthodologie utilisée afin de concevoir le système d'information en plus de faire la description des différents ensembles de données obtenus. L'identification des différentes problématiques liées à chacun de ces ensembles de données y est aussi faite. Le Chapitre 4 présente la conception du système d'information. Les différentes sections abordent les options de stockage, établissent les structures communes des ensembles de données décrits précédemment et explicitent la définition d'un schéma de données avant d'énumérer les différentes étapes nécessaires au montage du système d'information dans les technologies choisies. Le Chapitre 5 porte quant à lui sur l'exploitation des données, notamment sur les outils de peuplement automatisé de la base de données ainsi que sur des outils permettant de produire différentes analyses sur l'état du système de transport. Les capacités du système d'information développé à s'adapter à de nouveaux ensembles de données ainsi que sa résilience face à l'augmentation de sa taille y sont aussi abordées. Finalement, une conclusion résume l'ensemble du mémoire, en plus de présenter différentes perspectives qui découlent de la réalisation de ce projet.

CHAPITRE 2

REVUE DE LITTÉRATURE

Ce chapitre propose une revue de littérature scientifique et technique liée aux thèmes principaux abordés dans cet ouvrage. La section 2.1 aborde différentes études portant sur les nouvelles infrastructures et leurs effets sur l'ensemble des réseaux de transports. La section 2.2 porte quant à elle sur la gestion de grands ensembles de données dans les transports, les entrepôts de données ainsi que sur la quantification des impacts dans le cadre d'analyses de cycle de vie.

2.1 Nouvelles infrastructures et effets

L'implantation de nouvelles infrastructures de transport s'accompagne généralement de modifications des comportements des usagers, autant en ce qui a trait au mode de déplacement qu'aux itinéraires choisis. Dans le but de bien cerner les effets de l'érection d'une telle infrastructure, certains aspects théoriques seront d'abord abordés, suivis d'études de cas récentes.

2.1.1 Aspects théoriques

Paradoxe de Braess

Il importe tout d'abord de mentionner qu'il n'existe pas de consensus scientifique quant aux bénéfices véritables de l'augmentation de la capacité routière. Le mathématicien Dietrich Braess a été parmi les premiers à faire preuve de réserves quant aux bénéfices obtenus de la construction de nouvelles infrastructures dans un article de 1968 originalement publié en allemand (Braess *et al.*, 2005). Le paradoxe qui y est décrit fait état de potentielles augmentations des temps de parcours suite à la mise en place de nouveaux liens routiers. Bien qu'elle semble contre-intuitive, plusieurs travaux effectués depuis viennent appuyer la conclusion tirée par Braess.

Demande induite

Au-delà de la possible présence du paradoxe de Braess, de nombreuses études tendent à indiquer une relation directe entre l'ajout de capacité routière et une augmentation de la circulation automobile (Pfleiderer et Dieterich, 1995) (Goodwin, 1996) (Goodwin et Noland,

2003). Pfleiderer et Dieterich (1995) expliquent le phénomène par le concept du budget-temps, qui représente le temps moyen en transport d'une personne. Ce budget-temps restant constant, une augmentation de la capacité routière s'accompagnant d'une augmentation de la vitesse sur le réseau risque de favoriser un choix de domicile plus distant du lieu d'activité principal, venant ainsi augmenter le nombre de kilomètres parcourus. Une meilleure fluidité sur le réseau devrait toutefois s'accompagner, à court terme, de meilleures vitesses et ainsi engendrer une réduction globale de la quantité de carburant utilisée et permettre des diminutions de concentration de différents polluants atmosphériques (Hansen *et al.*, 1993). Toutefois, un transfert modal du transport en commun vers l'automobile associé à la réduction des temps de parcours automobiles et l'apparition de nouveaux déplacements auparavant non effectués peuvent réduire l'importance de ce bienfait environnemental (Banister, 2005). Les autres bénéfices des augmentations de capacité routière, en dépit de la circulation induite, sont principalement axés sur les gains de mobilité et d'accessibilité accompagnés par des réductions des temps de parcours (DeCorla-Souza et Cohen, 1999). Le lien direct entre l'augmentation de la capacité routière et une augmentation de la circulation automobile est toutefois remis en cause par certains travaux (Prakash *et al.*, 2001).

Dans l'ensemble, une majorité d'auteurs accepte la présence d'une demande induite suite à l'augmentation de la capacité. Certains travaux tentent d'ailleurs d'évaluer précisément l'augmentation de la demande relatif à l'augmentation de l'offre routière. Les valeurs d'élasticité (le rapport entre l'augmentation des kilomètres parcourus et l'augmentation de la capacité) sont toutefois très variables selon les auteurs et dans le temps (Hansen *et al.*, 1993) (Cervero, 2003) (Noland, 2001). Hansen *et al.* (1993) présentent des valeurs allant de 0,3 à 0,4 à court terme, et des valeurs de 0,4 à 0,6 après 16 ans. Cervero (2003) obtient des valeurs légèrement inférieures à court terme (0,238) et légèrement supérieures (0,637) à long terme. Finalement, Noland (2001) calcule des valeurs beaucoup plus élevées, avec des élasticités de 0,3 à 0,6 à court terme et de 0,7 à 1,0 à long terme. Dans tous les cas, les auteurs mentionnent que la demande induite devrait être prise en compte dans l'évaluation d'un projet. Dans quelques cas, des péages sont utilisés non seulement afin de financer la construction de nouvelles infrastructures, mais aussi pour contrôler l'augmentation du flot de véhicules (Piarc Technical Committee, 2012).

La complexité d'évaluer précisément l'intensité de la circulation induite par l'ajout de capacité routière amène certains problèmes. Cette difficulté nuit notamment à la mise en place de modèles de demande complets, ce qui rend difficile l'évaluation des impacts réels d'un projet sur les agents et l'économie (Laird *et al.*, 2005) (Metz, 2008). À cet effet, Metz (2008) cite que les gains en temps générés pour les usagers sont un des arguments majeurs utilisés pour justifier la mise en place de nouvelles infrastructures de transport, alors que très

peu d'études empiriques chiffrent ces gains.

Réponse des usagers

Un bon nombre d'études portent aussi sur la réaction des usagers face à des modifications de la capacité du réseau routier. La plus grande partie de ces travaux porte toutefois sur une réduction de la capacité. Néanmoins, le comportement des usagers face à ces modifications du réseau permet de déterminer certains facteurs affectant les choix d'itinéraires. Dans un premier temps, il est intéressant d'observer que les effets de réduction de capacité sont en général inférieurs à ce qui est anticipé (Goodwin *et al.*, 2002). De façon générale, il semble que les usagers préféreront changer leurs habitudes au sein d'un même mode, même dans des cas de réduction de capacité importante suite à des événements majeurs (Giuliano et Golob, 1998). À cet effet, Giuliano et Golob (1998) remarquent dans leur étude une diminution des effets de pointe et une meilleure répartition des déplacements dans le temps. Cette réaction se fera d'abord de façon amplifiée suite à une modification soudaine du réseau, puis mènera à une réévaluation des bénéfices obtenus par certains usagers, ce qui devrait conduire à une nouvelle situation d'équilibre (Clegg, 2007). Bien que chaque usager possède un budget-temps précis qui permet de calculer une valeur du temps moyenne (Brownstone et Small, 2005), un temps de déplacement constant de jour en jour reste un facteur d'importance pour un grand nombre d'usagers (Banister, 2005).

Données et infrastructures

Avant de procéder à une revue de différentes études de cas de mise en place d'infrastructures et des effets de celles-ci, il importe de mentionner deux constats présents dans une grande quantité d'analyses citées précédemment ainsi que dans nombre d'autres travaux. D'abord, plusieurs articles suggèrent le manque important d'études empiriques permettant de confirmer les différents modèles et conclusions développés (Bain, 2009) (Metz, 2008) (Short et Kopp, 2005) (Flyvbjerg *et al.*, 2006). Bain (2009) et Flyvbjerg *et al.* (2006) s'attardent surtout à la capacité de faire des prédictions justes à l'aide de modélisations. Metz (2008) analyse la situation d'un point de vue décisionnel des autorités en place, l'absence d'études empiriques rendant un choix éclairé entre différentes alternatives difficiles. Finalement, Short et Kopp (2005) recommandent un plus grand nombre d'évaluations post-ouverture afin d'améliorer la transparence des processus décisionnels et ainsi améliorer la perception du public face à ceux-ci.

Un autre constat effectué par nombre de publications concerne la difficulté d'obtenir des données, ou encore d'effectuer une bonne gestion de celles-ci. Quelques-un des problèmes

mentionnés incluent l'absence de documents en format électronique, le manque de documentation ou des difficultés d'accès aux données (Parthasarathi et Levinson, 2010). Dans d'autres cas, les auteurs s'inquiètent de la qualité des données qui sont rendues disponibles (Flyvbjerg *et al.*, 2006). De façon plus générale, plusieurs autres travaux mentionnent des inquiétudes par rapport aux données ou mentionnent une volonté d'accès à des données de meilleure qualité ou en plus grande quantité (Short et Kopp, 2005) (Flyvbjerg, 2005).

2.1.2 Études de cas

La quantité de travaux s'attardant aux impacts sur la circulation des nouvelles infrastructures est plutôt limitée. La majorité des travaux accomplis sont en général des études de pré-construction, alors que très peu portent sur les effets observés suite à la mise en service des projets. Une synthèse regroupant un très grand nombre de grands projets de transport a toutefois permis de déterminer que les études de pré faisabilité ont une tendance marquée à sous-évaluer les coûts de construction en plus de fortement mal évaluer l'utilisation réelle des structures une fois mises en service (Flyvbjerg *et al.*, 2003). En effet, plus de 50 % des projets étudiés montrent des surestimations ou des sous-estimations de plus de 20 % entre la circulation prévue et celle observée (Flyvbjerg *et al.*, 2006).

Il est à noter que l'ensemble des analyses effectuées par Flyvbjerg *et al.* (2003) porte uniquement sur les infrastructures construites et n'abordent pas les effets régionaux éventuels. Ces études en particulier sont encore moins nombreuses et sont plus difficiles à réaliser en raison de problèmes d'accès aux données, mais aussi en raison des difficultés de bien délimiter le secteur à étudier ainsi que de bien identifier les interactions entre les différents paramètres (Mackie et Preston, 1998). Certaines études plus poussées portant sur de nouvelles infrastructures et tentant d'évaluer des effets régionaux ont toutefois été identifiées et sont présentées ci-bas.

Pont de l'autoroute I-35W - Minneapolis

Un projet récent ayant mené à la rédaction de nombreuses publications est la reconstruction de l'autoroute I-35W à Minneapolis, Minnesota, aux États-Unis. Le pont d'origine traversant le fleuve Mississippi s'est effondré le 1er août 2007, retirant du réseau routier de la ville une artère autoroutière de 8 voies traversée par plus de 140 000 véhicules quotidiennement. L'effondrement du pont a eu des effets instantanés sur les usagers et leur choix en matière de transport. Sa reconstruction et sa remise en service un peu plus d'un an plus tard ont à nouveau perturbé les habitudes des usagers, permettant ainsi de faire des études sur un grand nombre des aspects théoriques énoncés plus tôt.

Études post-effondrement

Les premiers travaux effectués ont porté sur les conséquences de l'effondrement d'août 2007. Dans un premier temps, une modélisation des choix individuels a été accomplie. Celle-ci a permis de déterminer que les temps de déplacements anticipés ont été corrigés par les usagers après un certain nombre de jours, ce qui démontre qu'en cas de perturbation du réseau, une certaine période d'apprentissage est requise pour revenir à un point d'équilibre (He *et al.*, 2008), tel que proposé par Clegg (2007).

Le retrait du pont a aussi entraîné une période de pointe présentant un débit maximal moins important et une durée plus longue, même si le nombre de véhicules sur l'ensemble du réseau est resté plutôt constant. L'achalandage des transports en commun est resté plutôt constant lui aussi, ce qui semble confirmer la tendance des utilisateurs à modifier leurs comportements sans toutefois changer de mode de déplacement (Zhu *et al.*, 2009). La préparation des modèles et les constats en lien avec les modifications des comportements ont par la suite permis d'évaluer les coûts économiques quotidiens liés au retrait du pont. Ces coûts ont été évalués comme étant situés entre 70 000 et 200 000 $ par jour (Xie et Levinson, 2011).

Finalement, l'ensemble des mises à jour aux modèles régionaux utilisés qui ont suivi l'effondrement du pont de l'I-35W a été intégrée dans l'évaluation des coûts et bénéfices du projet de reconstruction du pont Lafayette, un autre pont de Minneapolis (Zhu et Levinson, 2008).

Études post-reconstruction

La reconstruction du pont a de nouveau modifié les comportements des usagers et a elle aussi fait l'objet de plusieurs recherches ciblées. Avant de procéder, une revue des pratiques et des données utilisées par d'autres projets a été faite, avec le constat que plusieurs expériences ne combinaient pas un assez grand nombre de données différentes et avaient une portée spatiale trop faible (Zhu et Levinson, 2010).

Parmi les études réalisées, un modèle de choix de pont a notamment été construit suite à la réouverture (Carrion et Levinson, 2012). Un sondage effectué auprès d'usagers quotidiens ainsi que la collecte de données GPS embarqués dans les véhicules de ceux-ci a permis de déterminer que les utilisateurs préfèrent les ponts les plus proches de leur lieu de domicile ou d'activité aux dépens des autres alternatives à mi-chemin. L'intérêt des usagers potentiels pour la nouvelle infrastructure était surtout axé sur le potentiel de gains en temps de parcours et sur la volonté d'obtenir une stabilité des temps de parcours quotidiens. Les utilisateurs qui ont fait le choix de se tourner vers le nouveau pont l'ont donc fait avec l'espoir d'améliorer leurs déplacements, alors que ceux qui l'ont ignoré associaient un coût élevé à la recherche d'alternatives à leur itinéraire habituel et démontraient un certain pessimisme par rapport

au potentiel de diminuer leur temps de déplacement.

Une évaluation d'un point de vue géographique a aussi été accomplie afin de déterminer les usagers ayant vu leurs temps de transport être réduit par la réouverture du pont, toujours à l'aide de GPS embarqués dans les véhicules. Les résultats de celle-ci ont aussi permis d'en apprendre plus sur les comportements des usagers suite à la mise en service du pont. Il y est notamment démontré que sur l'échantillon obtenu, les usagers n'effectuent pas de changement d'itinéraire, à moins que leurs gains en temps soient de plus de 10 minutes. Certaines mesures de mitigation avaient été mises en place pour pallier la réduction de capacité du réseau. Une fois celles-ci retirées, il apparaît aussi que le nombre de personnes ayant bénéficié de temps de parcours réduits suite à la construction du nouveau pont soit inférieur au nombre de personnes ayant connu une augmentation des temps de parcours. Géographiquement, les bénéfices sont concentrés dans les secteurs à proximité du pont, alors que les zones plus distantes présentent presque toutes des augmentations de temps de parcours pour leurs résidents (Zhu *et al.*, 2010).

Deux autres études confirment aussi certains des éléments avancés précédemment, notamment en ce qui a trait à la constance des habitudes des usagers ainsi que des choix basés sur les expériences passées. Les conclusions d'une de celles-ci sont que le nombre de déplacements en automobile a légèrement diminué suite à l'effondrement du pont en 2007, mais que la circulation suite à l'ouverture du nouveau pont était inférieure à la situation de référence. Ainsi certains déplacements qui étaient effectués avant la destruction du pont ne l'étaient plus suite à la mise en service de la nouvelle infrastructure (Danczyk *et al.*, 2010). La seconde étude est venue confirmer ces conclusions, en plus de définir géographiquement les déplacements soustraits, ceux-ci étant surtout situés à proximité du pont. (Zhu *et al.*, 2012).

Autres projets

De nombreux autres projets ont fait l'objet d'études quant à leurs effets sur le réseau, bien que très peu aient été étudiés en profondeur comme le pont de l'I-35W de Minneapolis. Néanmoins, plusieurs travaux abordent des aspects ignorés par les différentes études de Minneapolis qui sont d'intérêt afin de cibler précisément les effets des nouvelles infrastructures.

Autoroute M-40 - Madrid

Une première étude tente de cibler la variation d'accessibilité résultant de la mise en place de l'autoroute de contournement M-40 à Madrid (Gutierrez et Gómez, 1999). La saturation de la première autoroute de contournement M-30 a poussé les autorités espagnoles à construire une seconde voie de contournement de 1990 à 1996. Les résultats ont montré que les effets de la mise en place de cette nouvelle autoroute ont été surtout dans des secteurs à proximité de celle-ci. Les déplacements de contournement visant à éviter le centre de Madrid se sont

réaffectés sur l'autoroute M-40, l'autoroute M-30 devenant principalement utilisée pour des déplacements intraurbains. Par ailleurs, la mise en service de l'infrastructure n'a pas entraîné de réduction d'affluence sur le reste du réseau, alors que, dès son ouverture, l'autoroute M-40 a connu des problèmes de congestion en pointe. Les valeurs calculées pour les différents indicateurs de mobilité et d'accessibilité développés dans le cadre de l'étude se sont toutefois améliorées suite à la mise en service de l'autoroute.

Autoroute M-25 - Londres

L'autoroute de contournement M-25, à Londres, a aussi été étudiée, les travaux se concentrant cette fois sur les effets sur le développement économique régional (Linneker et Spence, 1996). Une relation négative entre l'accessibilité mesurée et les changements de niveau d'emploi a été obtenue. Ainsi, la mise en service s'est accompagnée d'effets négatifs sur le développement économique à proximité de la nouvelle autoroute. Les conclusions remettent en question les arguments quant aux effets bénéfiques des nouvelles infrastructures sur les milieux environnants. Pour expliquer la relation négative, les auteurs avancent que si la mise en service a amélioré l'accessibilité du secteur, cela a eu pour effet de le placer en concurrence avec les autres pôles régionaux, ce qui a nui aux entreprises locales se retrouvant désormais en concurrence avec des entreprises extérieures. Par ailleurs, comme dans le cas de Madrid, une congestion a été remarquée dès la mise en service pour l'autoroute M-25.

POPE - Royaume-Uni

Les autorités britanniques ont également développé une méthode standardisée d'évaluation des nouvelles infrastructures. Ces évaluations de projets suite à l'ouverture (la dénomination anglaise utilisée est Post-Opening Project Evaluation (POPE)) sont mandatées par la UK Highways Agency afin de mesurer l'achalandage précis sur une nouvelle infrastructure. Ces analyses sont toutefois concentrées sur les nouvelles routes et ne tiennent pas compte des effets à l'échelle régionale. Elles ont tout de même pu démontrer, dans certains cas, des gains de temps pour les déplacements automobiles suite à l'augmentation de capacité routière (Metz, 2008) (Highways Agency, 2006).

Autoroutes I-80 et I-880 - Oakland

Finalement, deux projets de reconstruction d'infrastructure à Oakland, en banlieue de San Francisco (Californie) ont aussi été étudiés. Les autoroutes I-80 et I-880 ont toutes deux été partiellement détruites par un tremblement de terre en 1989, et des études sur la remise en service de certains tronçons ont été complétées, basées sur des comptages routiers ainsi que des sondages en bord de route. Une première publication relate les effets de la reconstruction

d'un segment de route reliant l'autoroute I-80 à l'autoroute I-880. Les flux de véhicules qui y ont été calculés étaient en variation constante en fonction des jours, des mois et des années, mais une hausse générale a été remarquée suite à l'ouverture du tronçon étudié (Dahlgren, 1998).

Les réponses des usagers ont aussi été étudiées suite à la réouverture d'une section de l'autoroute I-880. L'analyse des réponses à des sondages en bord de route a permis d'évaluer les perceptions des usagers à propos de la nouvelle infrastructure et d'analyser les modifications de leurs habitudes. 41 % des répondants ont affirmé qu'ils auraient entrepris leurs déplacements à des moments de la journée différents sans la réouverture du segment étudié. Parmi les autres résultats d'intérêt, 17 % ont vu leur temps de déplacement diminuer suite à la réouverture de l'autoroute, 9 % de l'ensemble des répondants envisageaient de déménager plus loin de leur lieu d'activité principal, 11 % étaient désormais disposés à prendre un emploi situé à plus grande distance de leur lieu de résidence. Finalement, 3 % des répondants avancent qu'ils ne se seraient pas déplacés si l'autoroute n'avait pas été en service, confirmant ainsi la présence de circulation induite (Dahlgren, 2001). D'autres travaux portant sur la même section ont aussi permis de dénoter un transfert modal du transport en commun vers des déplacements automobiles, 7 % des répondants déclarant qu'ils auraient préféré le transport collectif à la voiture sans la mise en service de l'autoroute (Dahlgren et Station, 2003).

2.2 Gestion des données en transport et entrepôts de données

La littérature présentée dans la section précédente fait état de problèmes liés aux données, alors que la section qui suit fera état de différents travaux portant sur la gestion des données dans le domaine des transports ainsi que sur les entrepôts de données. Tout d'abord, les types de données associés aux transports seront abordés individuellement suivi d'une revue d'un ensemble de travaux portant sur l'intégration de ces données. Finalement, les outils permettant d'en faire l'intégration seront abordés.

2.2.1 Données en transports

Types de données

De nombreux travaux portent sur les données liées aux transports et à leur gestion. L'ensemble des données peut être ramené à quatre grandes familles, soit les données d'offre, de demande, de performance et d'impacts (Jack Faucet Associates, 1997).

Données d'offre

Jack Faucet Associates (1997) présente les données d'offre comme étant le potentiel offert aux éventuels utilisateurs. Dans le cas du réseau routier, des données comme une valeur de voie-km, le nombre total de kilomètres disponibles sur le réseau, la capacité d'un lien ou encore une numérisation complète du réseau feront partie de cette catégorie de données.

La nature même des données liées au transport implique la notion de positionnement dans l'espace. Pour cette raison, un aspect d'importance primordiale dans la gestion de ces données est d'en faire un traitement géographique. L'ensemble des pratiques consistant à gérer des ensembles de données de transport de façon géographique se nomme Système d'Information Géographique appliqué aux transports (SIG-T). La forme la plus courante de SIG-T est une représentation numérisée du réseau routier (Waters, 1999). Dans ces cas précis, l'ensemble des liens routiers est numérisé sous la forme de lignes et de points permettant de faire le lien entre les lignes. Ces numérisations des réseaux peuvent être conservées dans des Système de Gestion de Base de Données Relationnelle (SGBDR). La représentation du réseau devrait toujours se faire à l'intérieur d'un modèle de données général qui contient les paramètres des objets géographiques et non seulement leur représentation graphique (Dueker et Butler, 2000).

Les premiers efforts pour procéder à des numérisations ont été entamés dans les années 1960 dans le but de géolocaliser les données du recensement américain (Goodchild, 2000). Goodchild (2000) identifie trois représentations des réseaux, soit la vue cartographique, qui est la forme la plus classique de représentation d'objets localisés, la vue navigation, qui permet notamment de procéder à des calculs d'itinéraires sur des réseaux, ainsi que la vue comportementale, qui s'intéresse aux déplacements des objets sur le réseau.

Données de demande

Les données de demande consistent en l'utilisation qui est faite par les différents usagers du réseau disponible. Des informations comme le nombre de véhicules passant à un point précis d'évaluation du pic de congestion ou les comportements généraux des usagers font partie cette famille. La demande en transport peut aussi inclure des déplacements latents non réalisés. D'autres informations, comme des tracés Global Positioning System (GPS) de véhicules, peuvent aussi fournir des informations sur l'utilisation du réseau.

Une des formes les plus communes de données liées à la demande se présente sous la forme de données d'enquête de déplacement. Un très grand nombre de modèles et d'analyses basés sur ce type de donnée est présent dans la littérature, mais assez peu abordent la gestion de ces données. Certains travaux établissent des façons de les gérer dans des bases de données relationnelles (Shaw et Wang, 2000), ou encore font état de modèles et d'analyses basés sur

des traitements orientés-objet (Trépanier et Chapleau, 2001) (Frihida *et al.*, 2008).

Les données de comptage de véhicules issues de boucles de détection sur le réseau font aussi partie de la famille des données de demande observée. Les boucles de détection peuvent avoir des niveaux de résolution variables, certains n'accumulant des données que sur le nombre de véhicules ayant traversé le point d'analyse, alors que d'autres évaluent la vitesse ainsi que les types de véhicules (Bonsall et O'Flaherty, 1997). Bien que certains travaux soient faits dans le but de faire un traitement en temps réel des données accumulées (Chen et Petty, 2001), l'utilisation de données archivées est la plus fréquente. Un autre usage digne de mention des données de comptage est dans l'ajustement de l'utilisation de certains liens lors de la calibration de certains modèles de demande (Miller, 1999).

Finalement, l'utilisation de nouvelles technologies comme des GPS embarqués à bord des véhicules permet aussi d'obtenir un portrait de la demande réelle en transport d'une partie des usagers (Quiroga, 2000) (Barron *et al.*, 2004).

Données de performance

Les données de performance sont les données d'analyses issues de la relation entre les données d'offre et de demande. Des évaluations de performance du réseau en fonction des temps de parcours ont été faites, notamment à l'aide de GPS (Barron *et al.*, 2004), de systèmes d'identification des plaques minéralogiques (Quiroga, 2000) ou encore à l'aide de boucles de détection (Kwon *et al.*, 2000). En outre, les données de performance sont utilisées dans certains cas afin d'effectuer la gestion de la circulation à l'aide de systèmes intégrés (Fu *et al.*, 2006). Les données de performance pour un réseau routier devraient idéalement comporter des informations sur les vitesses, les flots de véhicules, la densité et le débit sur les différents liens, ainsi que les types de véhicules, les incidents et restrictions sur le réseau, les péages et autres frais d'utilisation, de même que la description de l'offre des différents liens routiers. L'ensemble de ces données devrait aussi faire l'objet de mises à jour fréquentes (Pisarski, 1997).

Il est possible de définir une série d'étapes permettant de compléter la collecte des données, l'analyse et la prise de décision découlant de l'analyse (Dahlgren *et al.*, 2001). Un tel cycle doit être constamment repris, compte tenu des modifications qu'engendrent les prises de décisions des différents agents. Dahlgren *et al.* (2001) définissent ces étapes comme suit : des conditions de circulation sont mesurées par des outils de terrain qui les communiquent à un organisme central. Celui-ci les analyse pour faire des études de performance, puis les données sont archivées, les résultats des analyses disséminées, ce qui permet la prise de décisions par les organisations ou personnes responsables, qui peuvent mener à des modifications du réseau et donc à de nouvelles conditions de circulation, qui doivent être étudiées. Des problèmes peuvent

toutefois survenir si une étape du cycle n'est pas accomplie correctement. Une représentation graphique du cycle peut être observée à la Figure 2.1.

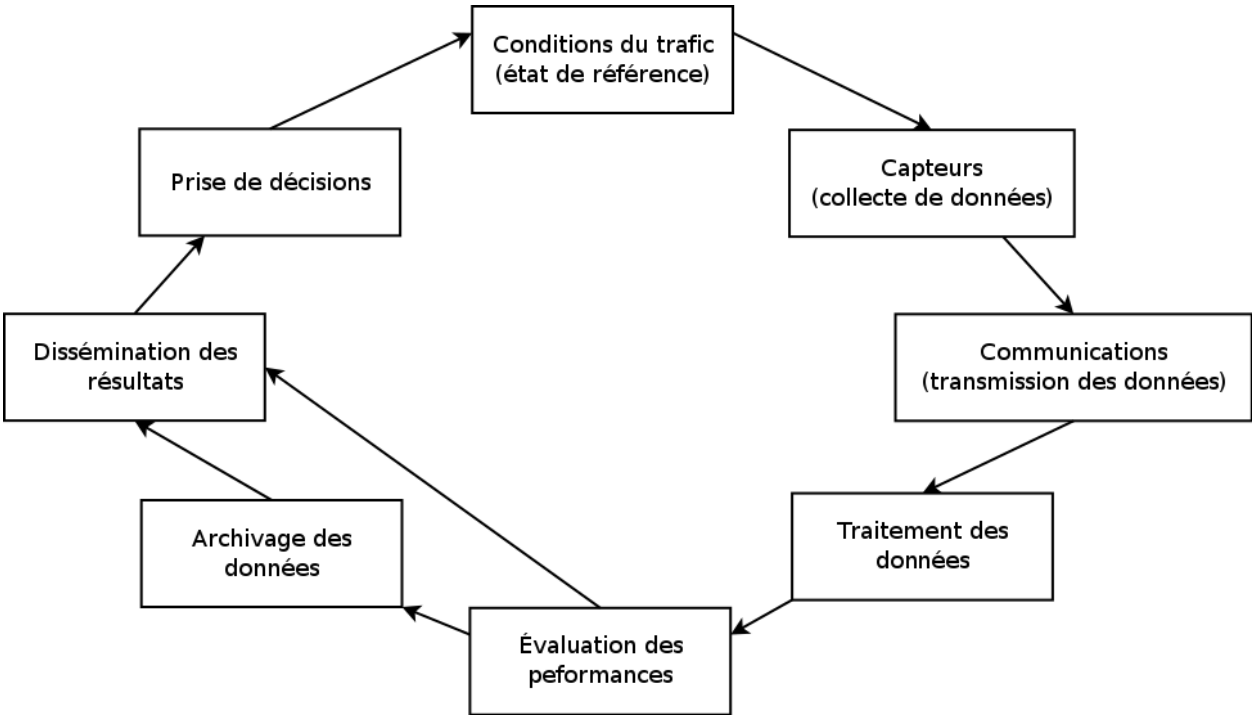


Figure 2.1: Cycle de collecte de données et de prise de décision (traduit de Dahlgren *et al.* (2001))

Données d'impact

Finalement, les données d'impact sont celles faisant état des externalités causées par l'utilisation des réseaux de transport. Dans le cas des réseaux routiers, ces impacts pourront être de nature environnementale, économique ou sociale (Huang, 2003). En ce qui concerne les impacts environnementaux Bonsall et O'Flaherty (1997) recommandent de faire des analyses avant et après toute modification au réseau afin d'en mesurer les effets.

Intégration des données

L'intégration de l'ensemble des données liées à un objet d'étude présente plusieurs avantages. Notamment, cette intégration devrait permettre à long terme d'économiser des ressources en permettant d'éviter que les mêmes solutions à certains problèmes techniques doivent être développées par l'ensemble des utilisateurs des données (Ziliaskopoulos et Waller,

2000). En outre, une bonne intégration favorise le partage des données et devrait permettre d'améliorer la réutilisabilité des données, ce qui en augmente la valeur intrinsèque (Etches *et al.*, 1998) (Schofer, 2007). Schofer (2007) remarque aussi que l'organisation des données et un schéma organisé devraient permettre d'obtenir des réponses rapides et ainsi améliorer et accélérer les processus de prise de décision. Une représentation plus précise et ordonnée permet la création de bases de données spécifiques à certaines applications à partir de données élémentaires, ce qui évite le dédoublement des données pour chaque application (Dueker et Butler, 2000). Une bonne organisation des données permettra également d'utiliser plus efficacement des méthodes de data mining ou encore d'utiliser des algorithmes d'apprentissage automatique (Hulten *et al.*, 2001) (Pyle, 1999) (Chen *et al.*, 1996).

L'intégration de données hétérogènes nécessite toutefois le développement de certains algorithmes et règles de conversions afin de permettre que des données conservées dans des projections, échelles ou modèles différents puissent être ramenées à des structures communes (Thill, 2000). Thill (2000) avance que la création d'un système de gestion de données unique pour faire des analyses multi thématiques représente le défi le plus important, et suggère qu'un modèle orienté-objet représente la meilleure solution pour y arriver.

Certains travaux font état des différentes étapes et cheminements accomplis afin de procéder à l'intégration de grands ensembles de données de types et de sources variés (Huang, 2003) (Valsecchi *et al.*, 1999). Huang (2003) mentionne le besoin, pour l'utilisation de modèles, d'avoir des données à des échelles spécifiques. Ainsi, pour certaines applications, il est nécessaire d'avoir des données agrégées à des niveaux supérieurs ou désagrégées à des niveaux inférieurs. Cinq tâches principales sont énoncées dans le travail d'intégration : la standardisation des ensembles de données, la mise en place d'une interface commune pour les ensembles de données, le référencement spatial et temporel, l'agrégation ou la désagrégation et finalement le stockage des données.

Valsecchi *et al.* (1999) présente un système d'intégration avec deux objectifs principaux, soit améliorer les capacités graphiques et les fonctions de base de données d'un système de gestion de données de circulation et développer des outils d'analyse temporelle animés. Pour y arriver, une base de données historiques des données de circulation est proposée afin d'étudier les conditions du trafic dans le temps et l'espace, le tout en complément d'analyses en temps réel.

2.2.2 Stockage des données

La nature des données liées aux analyses en transport ayant été énoncée, il est essentiel d'aborder les questions liées au stockage de ces données et aux problématiques d'organisation de l'information. Des ouvrages faisant la référence de bonnes pratiques dans la conception

de schémas de bases de données ainsi que d'autres abordant plus spécifiquement le chaos inhérent au stockage de grandes quantités de données permettent de diriger le développement du système d'information.

Les travaux abordant spécifiquement les méthodes utilisées afin de faire le stockage et l'utilisation des données étant limités dans les travaux énoncés précédemment, il est pertinent d'étendre la littérature à des références couvrant principalement les bonnes pratiques liées à ce stockage afin de pouvoir les appliquer au projet en cours. Elmasri et Navathe (2011) couvrent dans leur ouvrage la plupart des pratiques modernes en ce qui a trait à la modélisation et à la conception de schéma de données. Des conceptualisations incluant notamment des schémas entité-association ainsi que l'approche orientée-objet y sont décrites en détail, offrant ainsi une base théorique sur laquelle peut se baser le développement d'un système d'information. Les capacités des SGBDR modernes y sont aussi abordées en détails permettant à terme de baser le développement logiciel qui devra être accompli sur des technologies et des méthodes de travail modernes.

Si les bonnes pratiques de conception de bases de données sont utiles dans le développement du système d'information, il est pertinent de les jumeler à d'autres méthodes de travail permettant notamment la définition de schémas qui pourront évoluer lorsque de nouvelles données viendront se greffer au projet. Ambler et Sadalage (2006) et Ambler (2012) établissent les grandes lignes que devraient prendre de telles structures, en plus d'énoncer certains grands principes devant les accompagner. S'inspirant de la méthode de développement Agile, l'emphase est mise sur la fonctionnalité, la compartimentation des méthodes développées et sur des améliorations simples et fréquentes. En outre, il y est suggéré que la base de données devrait être construite tout au long du projet et non dans une phase de design préalable qui ne pourrait anticiper tous les besoins. Finalement, des façons de faire afin de gérer le changement constant et les données archivées sont avancées.

Les structures définies et les informations qui y sont stockées sont toutefois susceptibles de devenir désorganisées, un phénomène qui peut être favorisé par la présence de structures changeantes. Olsen Jr (1999) définit cette situation comme étant un chaos inhérent engendré par la diversité de l'information, des collaborateurs et des plates-formes utilisées. Un cycle d'utilisation de l'information est aussi défini, avec trois grandes étapes, soit le chaos, l'organisation et finalement l'exploitation de l'information. La solution principale avancée pour pallier ces problèmes engendrés par le chaos est d'effectuer des modélisations basées sur l'état naturel de l'information. À cet effet, Olsen Jr (1999) mentionne toutefois que le monde modélisé représente toujours une abstraction idéalisée d'un monde réel qui ne l'est pas.

Brackett (1996) discute aussi des problématiques liées à l'entreposage des données, et avance que l'intégration de l'information peut mener à des redondances et à un chaos dans

l'ensemble des données. À ce sujet, Brackett (1996) propose plusieurs pistes pour éviter le désordre lié à l'accumulation de quantités toujours plus grandes de données. Le besoin de documenter l'organisation et la structure des données afin que les ajouts par des utilisateurs différents conservent la signification qui était définie à l'origine est aussi mentionné. En accord avec plusieurs travaux cités dans la section sur l'intégration des données, Brackett (1996) propose d'utiliser un entrepôt de données unique, même dans le cas des données hétérogènes, la multiplication des entrepôts engendrant des redondances inutiles et inefficaces. Finalement, les risques associés à une perte de contrôle des données sont détaillés. Parmi ces risques, la multiplication des formats de données combinée à une augmentation exponentielle de la quantité de données disponible est présentée comme étant la principale source de chaos dans la gestion des données.

CHAPITRE 3

MÉTHODOLOGIE GÉNÉRALE

Ce chapitre a pour objectif de faire une description du contexte régional dans lequel le nouveau pont s'insère, de présenter les mandats associés au programme de suivi, de présenter une méthodologie pour réaliser le projet, en plus d'étudier les données qui peupleront le système d'information. Afin de bien répondre aux besoins d'analyses, il est nécessaire de situer le contexte géographique et de définir les travaux qui devront être accomplis en se basant sur le système développé. Tout d'abord, l'espace dans lequel la plupart des analyses à effectuer se situeront sera abordé. Par la suite, les mandats d'analyse eux-mêmes seront détaillés, ainsi que la méthodologie utilisée pour développer un système d'information répondant aux besoins. Finalement, chacun des types de données, soit des données de qualité de l'air, de météo et de circulation, ainsi que les formats dans lesquels elles sont fournies, seront détaillées et analysées afin d'y identifier des problématiques devant être résolues pour que le système d'information résultant présente des valeurs normalisées et cohérentes.

3.1 Contexte régional

Le nouveau pont de l'autoroute 25 s'insère dans un réseau routier régional déjà développé. En effet, six ponts font déjà le lien entre Laval et l'île de Montréal, en plus de deux autres ponts liant directement l'île de Montréal avec la rive nord du fleuve Saint-Laurent. Le réseau routier supérieur est aussi fortement développé dans le secteur à l'étude, avec trois autoroutes dans l'axe est-ouest et deux dans l'axe nord-sud. L'achèvement du nouveau pont vient compléter l'axe de l'autoroute 25, dont la construction a été entreprise par la mise en place des premières phases en 1966 et 1967, notamment avec la réalisation du pont-tunnel Louis-Hippolyte-Lafontaine reliant l'île de Montréal à la Rive-Sud.

Le décret 1243-2005 du gouvernement du Québec, qui a autorisé le processus de construction du pont, spécifie sa géométrie ainsi qu'un ensemble de contraintes entourant l'administration du pont, notamment concernant le partenariat public-privé mis en place pour la construction ainsi que les règles entourant les tarifs imposés aux utilisateurs du pont. À cet effet, le décret spécifie que l'autoroute elle-même ne doit comporter qu'au maximum quatre voies, et le pont six voies de circulation.

La Figure 3.1 fait état de la position des infrastructures importantes directement connectées au pont Olivier-Charbonneau ainsi que les autres structures routières offrant des alter-

natives, en plus de localiser le segment de l'autoroute 25 ayant fait l'objet du développement couvert par le décret.

Figure 3.1: Principales infrastructures de circulation du secteur à l'étude

3.2 Méthodologie

Différentes philosophies de développement logiciel peuvent servir dans le processus de planification du système d'information. Si la plupart de ces philosophies impliquent une planification minutieuse incluant une conception complète avant la mise en place du système, d'autres, comme la méthode de développement Agile, mettent l'emphase sur une mise en service rapide. La méthode Agile propose en effet une gestion de développement itérative incluant une implantation dès les premières étapes, suivie de nombreuses mises à jour en fonctions des besoins qui s'ajoutent. La méthode permet en outre d'intégrer des rétroactions issues des différents problèmes identifiés dans les premières itérations. Elle met aussi l'accent sur la possibilité d'effectuer des modifications rapides lorsque jugées appropriées et la plus

grande réutilisation possible des solutions développées (Martin, 2003).

Dans le cadre de ce projet, l'ensemble des travaux peut être ramené à cinq étapes principales, soit la description des fichiers et des formats de données obtenus, l'identification de toutes les sources d'erreurs potentielles et incohérences dans ces fichiers, l'identification de structures communes, l'élaboration d'un schéma de données et finalement l'exploitation de ces données stockées. Un diagramme résumant les travaux et faisant état des diverses rétroactions peut être observé à la Figure 3.2. Les couleurs utilisées reprennent le cycle proposé par Olsen Jr (1999), le bleu représentant le chaos de l'information, le jaune l'organisation de l'information et le vert la phase d'exploitation.

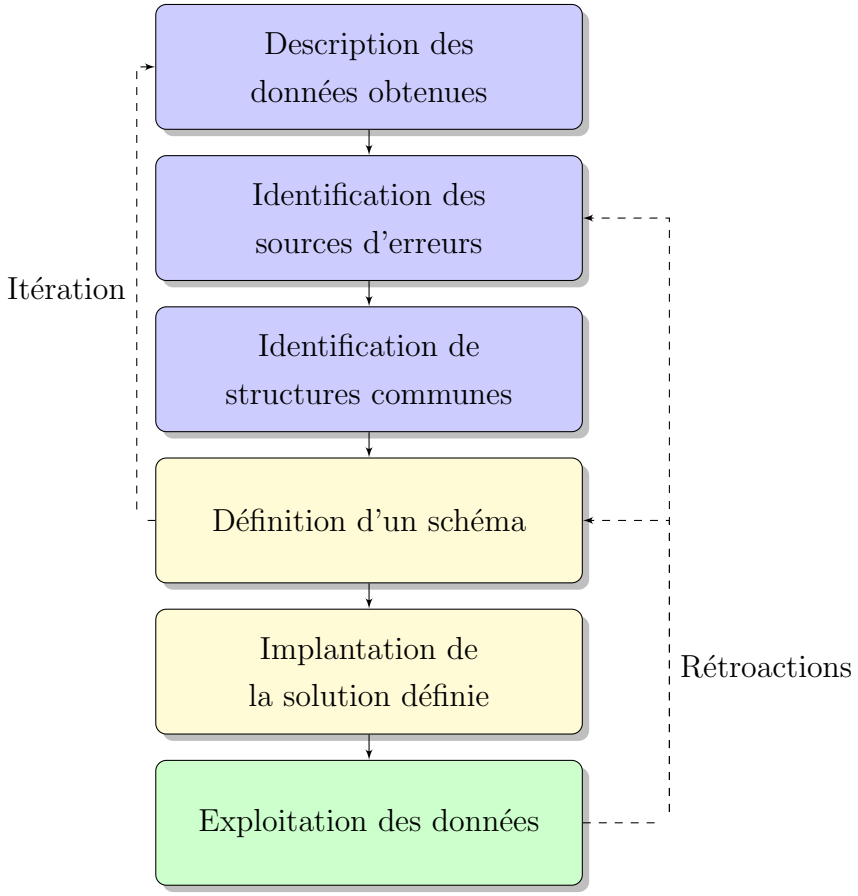


Figure 3.2: Diagramme de la méthodologie appliquée

La nature itérative du processus de développement selon la méthode Agile permet d'obtenir un système d'information de base très rapidement. Les étapes énoncées précédemment devraient donc être exécutées de façon itérative et être accomplies à chaque fois qu'un nouvel ensemble de données est rendu disponible. Les phases d'exploitation de données devraient quant à elles être accompagnées de rétroactions permettant de faire des ajustements struc-

turels ou sur les données elles-mêmes si des lacunes sont détectées.

Il importe alors de développer une solution modulaire, qui fait en sorte que des modifications jugées nécessaires puissent être intégrées sans toutefois avoir à reprendre les processus d'intégration des données déjà ajoutées au système d'information. Les différentes étapes du processus de développement sont énoncées subséquemment, en mettant l'emphase sur l'application des principes de développement itératifs.

3.2.1 Description des données obtenues

La description des données obtenues consiste à identifier les types de formats dans lesquels celles-ci se présentent. Chaque ensemble et format devra faire l'objet de description complète avant de pouvoir être intégré au système d'information. En outre, la description des données devrait permettre d'identifier des paradigmes, qui contrôleront éventuellement la structure du schéma de données qui sera développé. Ainsi, la nature itérative de la méthode Agile permettra de procéder en séquence et d'obtenir rapidement un accès aux premiers ensembles de données et de procéder à l'intégration selon les besoins d'analyse.

3.2.2 Identification des sources d'erreur

Chaque ensemble de données se présente avec un ensemble de défauts et de potentielles erreurs, autant dans la structure même des fichiers que dans la codification de l'information. L'identification de ces sources d'erreurs permet d'assurer la cohérence des données qui seront stockées dans la base de données en plus de simplifier les travaux aux étapes d'analyses. Les ensembles de données étant en général de grandes tailles, il est difficile d'identifier toutes les problématiques immédiatement. L'utilisation d'un processus d'implantation itératif et de rétroaction devrait permettre de pouvoir gérer dès leur identification les erreurs présentes, de les réparer lorsque possible et ainsi d'assurer un contrôle de qualité sur le système d'information.

3.2.3 Identification de structures communes

L'identification de structures communes à tous les ensembles de données doit se faire en raison des sources de données multiples et de la nature divergente des contenus. L'objectif d'un système d'information est de codifier et structurer des informations qui ne sont pas directement compatibles à l'état brut. La définition d'un paradigme pour l'ensemble des informations obtenues permet de générer un ensemble de structures et de normes qui permettent une organisation rendant aisément accessible l'ensemble de l'information acquise.

3.2.4 Définition d'un schéma

Les structures communes identifiées précédemment sont la base du schéma de données à développer. Celui-ci est essentiel dans le but de normaliser l'ensemble des informations afin de répondre aux besoins d'analyse. La normalisation permet de fournir une définition claire des différents objets simples en présence et à terme de permettre l'élargissement du cadre à d'autres types de données. En outre, combinée à des technologies appropriées, elle permet de définir des structures de validation de l'intégrité des données et d'éviter des incohérences telles que des informations redondantes ou contradictoires.

Le développement itératif fait en sorte que le schéma est lui-même en constante évolution suite à l'ajout de nouveaux ensembles de données. Les rétroactions permettent quant à elle d'ajuster la structure utilisée et encore une fois de simplifier les processus d'exploitation des informations.

3.2.5 Implantation de la solution définie

L'implantation de la solution définie consiste à formellement créer la structure de données définie précédemment dans des solutions logicielles choisies. Ces dernières doivent en outre répondre à un ensemble de besoins énoncés préalablement. Cette étape inclut l'insertion des données dans la structure définie ainsi que la gestion des différents problèmes identifiés, s'il y a lieu. Suite à l'implantation, un système d'information clairement défini est disponible et l'information est prête à être exploitée.

3.2.6 Mise en valeur du système de gestion de données

La dernière étape consiste à faire usage du système développé dans les phases précédentes. Elle doit notamment inclure des outils de visualisation de l'information et de production de résultats dans un format défini afin de pouvoir fournir des éléments d'analyse. Les difficultés d'exploitation ou d'éventuels problèmes structurels devraient permettre de revenir aux étapes précédentes et de faire des ajustements afin de simplifier et de rendre efficaces les processus d'accès aux données en présentant ces dernières sous une forme appropriée.

3.3 Données de qualité de l'air

Les données de qualité de l'air sont fournies par deux sources, soit le Ministère des Transports du Québec (MTQ) et le Réseau de Surveillance de la Qualité de l'Air de la ville de Montréal (RSQA), qui offrent des données respectivement pour six et trois stations. Dans les

deux cas, l'ensemble des stations accumule des mesures pour des molécules et particules, alors qu'un sous-ensemble présente aussi des données de Composés Organiques Volatils (COV).

3.3.1 MTQ

Le MTQ fournit des données pour un ensemble de six stations situées à proximité de l'axe de l'autoroute 25, deux de celles-ci offrant aussi des enregistrements concernant les COV. L'ensemble de ces stations est présenté à la Figure 3.3.

Données de molécules et particules

Les données pour les molécules et les particules [1] des stations du MTQ sont fournies à deux niveaux de résolution, soit des données présentant des lectures à chaque minute ou agrégées à l'heure, dans des bases de données du logiciel Microsoft Access. Les analyses ayant à être complétées se basant uniquement sur des mesures horaires, le choix a été fait d'ignorer les données avec des intervalles d'une minute et d'utiliser les données horaires.

Les données prennent la forme d'une table de données pour chaque couple de station et paramètre entre le 31 mars 2010 et le 21 mai 2011, pour un total de 46 tables. Le Tableau 3.3 fait état de l'ensemble des paramètres étudiés et du nombre d'entrées pour chacun de ceux-ci pour toutes les stations. La forme que prend chacune des tables dans le fichier d'origine est présentée à la Figure 3.4 pour le NO_2 à la station S2.

Date	NO2
12/11/2010 13:00	
12/11/2010 14:00	
12/11/2010 15:00	
12/11/2010 16:00	
12/11/2010 17:00	
12/11/2010 18:00	
12/11/2010 19:00	
12/11/2010 20:00	
12/11/2010 21:00	
12/11/2010 22:00	
12/11/2010 23:00	
12/11/2010 24:00	
13/11/2010 01:00	

Figure 3.4: Exemple de données du fichier fourni par le MTQ

1. Les COV, mentionnés plus tard, sont aussi des molécules, mais afin de faire la distinction entre les deux types de données, cette nomenclature est utilisée.

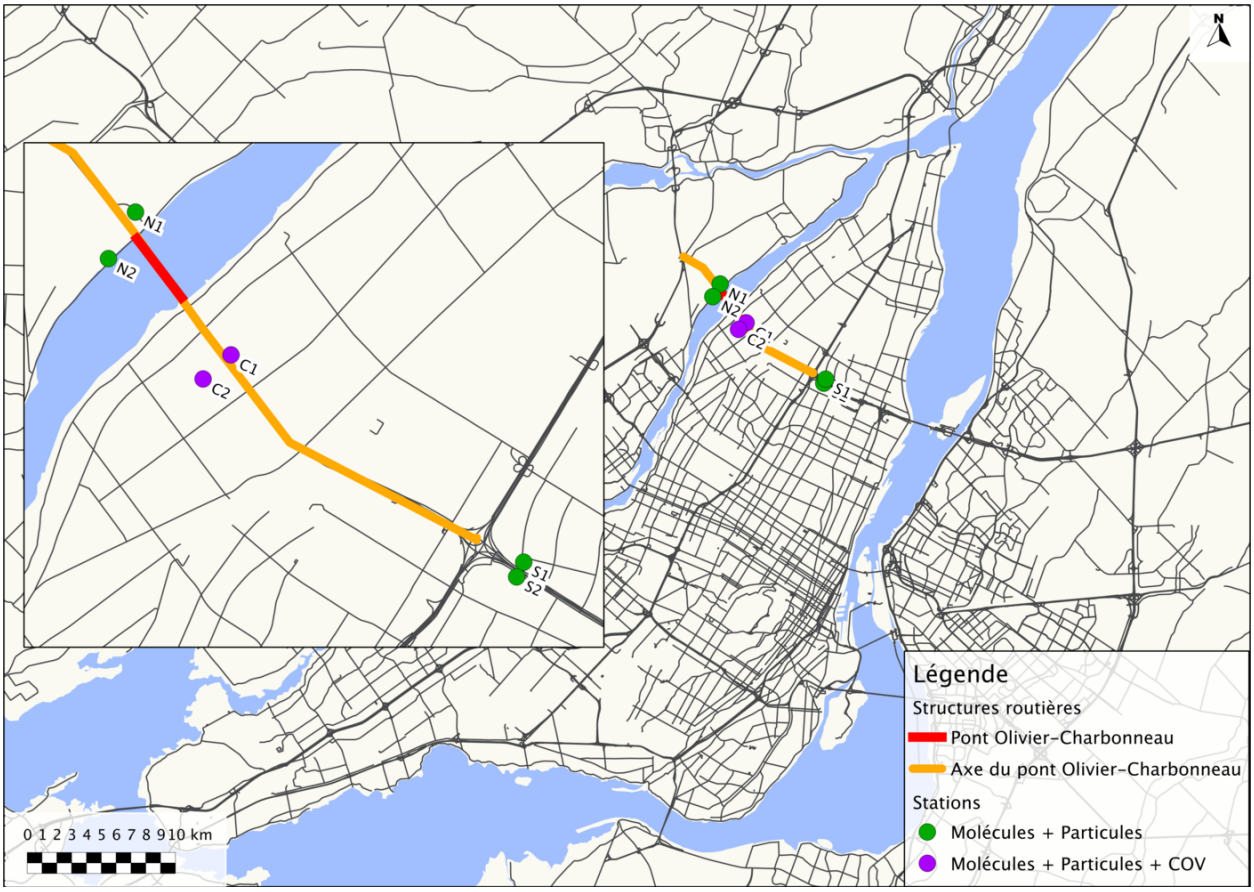


Figure 3.3: Stations de qualité de l'air du MTQ

Le format dans lequel les données sont fournies présente plusieurs éléments pouvant poser des problèmes d'utilisation des données et qui font en sorte que chacune des entrées devra potentiellement subir un traitement individuel avant de pouvoir être utilisée.

Tout d'abord, chacune des stations présente trois codes d'identification différents. Les codes numériques fournis dans les tables de données sont des codes à cinq chiffres, alors que la véritable nomenclature des stations utilise un code à deux caractères qui indique la position de la station, soit N1 et N2 pour les deux stations les plus au nord, C1 et C2 pour les deux stations centrales et S1 et S2 pour les deux stations les plus au sud. Les stations se voient toutes attribuer un nom indiquant l'artère routière la plus proche, et possèdent de plus un autre code numérique d'identification. L'ensemble de ces informations pour chacune des six stations est présenté au Tableau 3.1.

Tableau 3.1: Nomenclature des stations

Nom de la station	Code des tables Access	Code réel	Code supplémentaire
Lévesque	06209	N2	52
Roger Lortie	06208	N1	51
Perras	06057	C2	50
Autoroute 25	06056	C1	49
L-H Lafontaine	06005	S2	47
Châteauneuf	06006	S1	48

La structure des tables Access présente aussi un défi de gestion. Le nom de la colonne faisant l'accumulation des données étant variable en fonction du paramètre étudié, tout traitement automatisé devra être adapté de manière à utiliser la seconde colonne sans toutefois en spécifier formellement le nom. Le champ "Date" de chacun des enregistrements présente quant à lui l'horodatage de la fin d'un intervalle de mesure. Ces intervalles étant toujours d'une heure, l'horodatage fourni présente la mesure moyenne sur l'ensemble de l'heure précédente (ainsi, 12/11/2010 13:00 représente la mesure sur l'intervalle de 12/11/2010 12:01 à 12/11/2010 13:00). Par ailleurs, les unités pour les mesures sont variables en fonction des paramètres étudiés. Cette variabilité des unités de mesure fait en sorte que certaines entrées doivent potentiellement être modifiées afin d'être stockées dans une unité appropriée. L'ensemble des paramètres et des unités de mesure associées est présenté au Tableau 3.2

Un autre problème associé aux données de qualité de l'air du MTQ est que celles-ci sont toujours présentées à l'heure normale (UTC-5). Si la plupart des analyses ne portant exclusivement que sur les données de qualité de l'air peuvent être réalisées sans l'utilisation de l'heure civile, la mise en relation avec d'autres ensembles de données peut causer une incompatibilité. Puisqu'il sera vraisemblablement nécessaire de procéder à des analyses croisées de qualité de l'air avec d'autres paramètres, comme les débits de circulation par exemple, l'ensemble des mesures effectuées lors de la période d'utilisation de l'heure avancée (UTC-4) devra être ajusté afin de refléter le changement d'heure afin d'assurer la cohérence entre les différents types de données.

Tableau 3.2: Types de données obtenues par les différentes stations de qualité de l'air

Symbole	Nom complet	Unité
NO_2	Dioxyde d'azote	$\mu g/m^3$
NO	Monoxyde d'azote	$\mu g/m^3$
NO_x	Oxydes d'azote	$\mu g/m^3$
O_3	Ozone	$\mu g/m^3$
CO	Monoxyde de Carbone	$0.1 * \mu g/m^3$
SO_2	Dioxyde de soufre	$\mu g/m^3$
PM_{10}	Particules	$\mu g/m^3$
PM_{25}	Particules fines [2]	$\mu g/m^3$
PST	Particules en suspension	$\mu g/m^3$

Finalement, un problème de gestion plus important est associé aux données moléculaires et de particules du MTQ. En effet, pour l'ensemble des données, les enregistrements pour un certain nombre d'heures sont manquants. L'observation des mesures prises à des intervalles d'une minute révèle que les stations sont toujours en fonction pour ces heures manquantes. Cette situation s'explique par le fait que les données sur une heure sont une moyenne sur l'ensemble des données à la minute, mais que cette moyenne est jugée valide uniquement si 75 % des entrées à la minute sont présentes et valides.

Les heures pour lesquelles ce critère de qualité n'est pas atteint ne sont pas présentes dans la base de données horaire fournie. Or, l'absence complète de ces enregistrements présente un problème de gestion étant donné qu'il est essentiel de pouvoir déterminer la fiabilité d'une station et sa performance dans le temps. Par ailleurs, le calcul de moyennes mobiles, une méthode d'analyse très fréquente dans le cas des données de qualité de l'air, est rendu plus difficile par l'absence de ces données. La solution la plus simple pour pallier ce problème est de procéder à l'ajout de valeurs nulles et non zéro pour les heures manquantes afin d'illustrer l'invalidité de la mesure faite sur une heure en particulier. Ces valeurs nulles ne seront pas prises en compte lors du calcul des moyennes mobiles.

Les figures 3.5, 3.6, 3.7, 3.8, 3.9, 3.10, 3.11, 3.12 et 3.13 offrent une représentation graphique pour les années 2010 et 2011 de la répartition temporelle pour chacune des stations des mesures obtenues dans le cadre du premier envoi de données. Les lignes horizontales y représentent les stations dans le même ordre que celui utilisé au Tableau 3.1, les pixels

2. Les données pour les particules fines peuvent être fournies selon deux méthodes de collecte, soit GRIMM et FDMS, en fonction des stations. Les stations du MTQ n'utilisent que la méthode GRIMM, alors que les stations du RSQA utilisent principalement la méthode FDMS.

individuels la valeur associée à un jour et les bandes grises verticales les changements de mois. Les couleurs de chacune des cellules présentent le nombre de mesures obtenues pour une journée, les couleurs près du vert indiquant un nombre s'approchant de 24, les couleurs près du rouge indiquant des valeurs près de 0 alors que l'absence de mesures est représentée par la couleur noire. L'étude des figures permet de confirmer visuellement quels polluants sont étudiés par chacune des stations. Dans tous les cas, les graphiques permettent aussi de voir que les données manquantes pour chaque station tendent à apparaître sur de longues séquences, alors que les discontinuités ponctuelles illustrées par des couleurs dans les tons de jaune et de rouge sont beaucoup plus rares. À noter que l'absence de données après le mois de mai 2011 est attendue, le présent mémoire ne s'attardant qu'aux données de la période de référence qui précède l'ouverture du pont.

Tableau 3.3: Nombre de données disponibles par station et par paramètre étudié pour les stations de qualité de l'air du MTQ

Stations	NO_2	NO	NO_x	O_3	PM_{10} (GRIMM)	PM_{25} (GRIMM)	PST (GRIMM)	CO	SO_2	PM_{25} (FDMS)
N2 - Lévesque	9 614	9 614	9 614	9 713	9 966	9 966	9 966	0	0	0
N1 - Roger Lortie	5 274	5 274	5 274	9 990	9 748	9 760	9 748	0	0	0
C2 - Perras	9 018	9 018	9 018	9 813	8 245	8 245	8 245	9 377	6 396	0
C1 - Autoroute 25	9 388	9 388	9 388	9 747	9 920	9 920	9 920	9 412	9 718	0
S2 - L-H Lafontaine	4 224	4 224	4 224	6 968	6 937	6 941	6 937	0	0	0
S1 - Châteauneuf	8 547	8 547	8 547	9 786	8 651	8 661	8 184	0	0	0

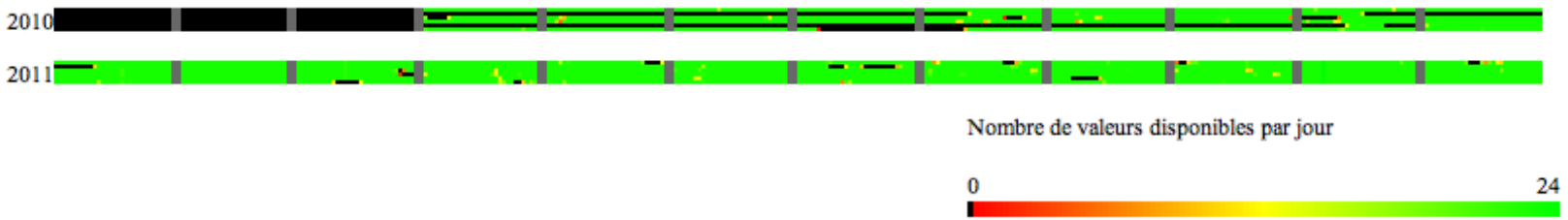


Figure 3.5: Couverture temporelle des données fournies par le MTQ pour le dioxyde d'azote

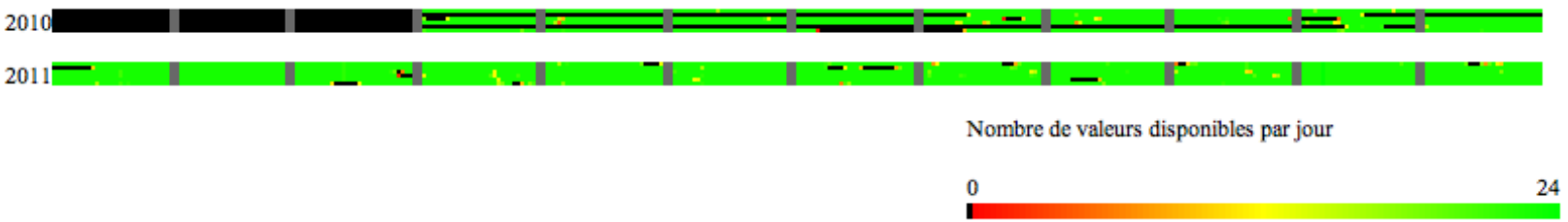


Figure 3.6: Couverture temporelle des données fournies par le MTQ pour le monoxyde d'azote

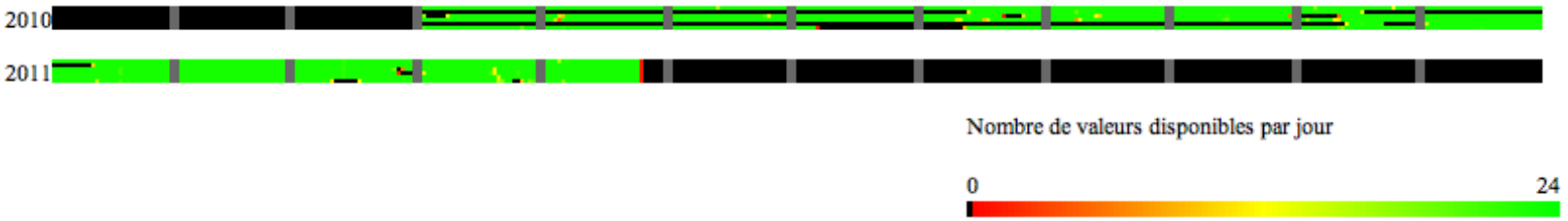


Figure 3.7: Couverture temporelle des données fournies par le MTQ pour les oxydes d'azote

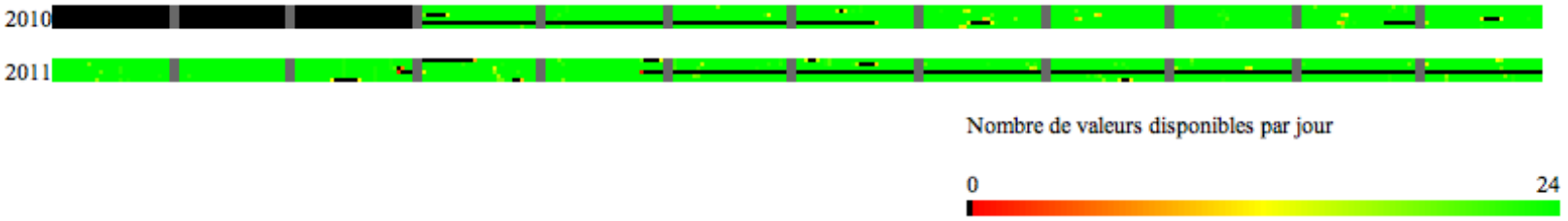


Figure 3.8: Couverture temporelle des données fournies par le MTQ pour l'ozone

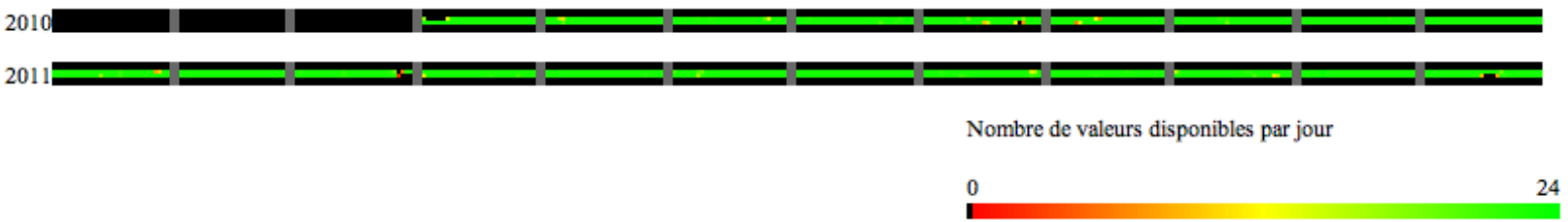


Figure 3.9: Couverture temporelle des données fournies par le MTQ pour le monoxyde de carbone

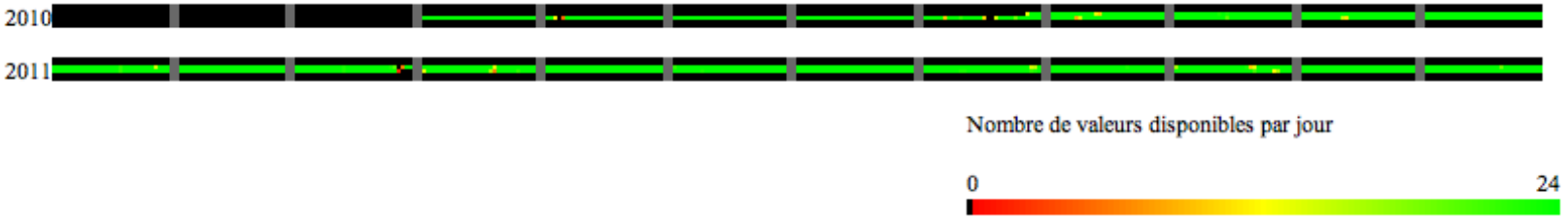


Figure 3.10: Couverture temporelle des données fournies par le MTQ pour le dioxyde de soufre

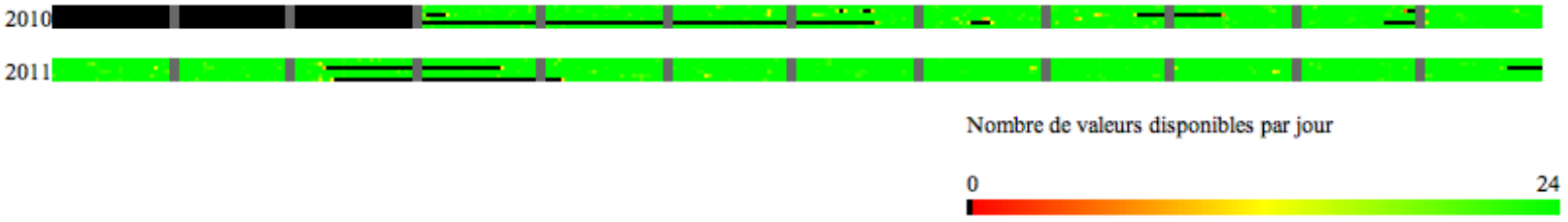


Figure 3.11: Couverture temporelle des données fournies par le MTQ pour les particules

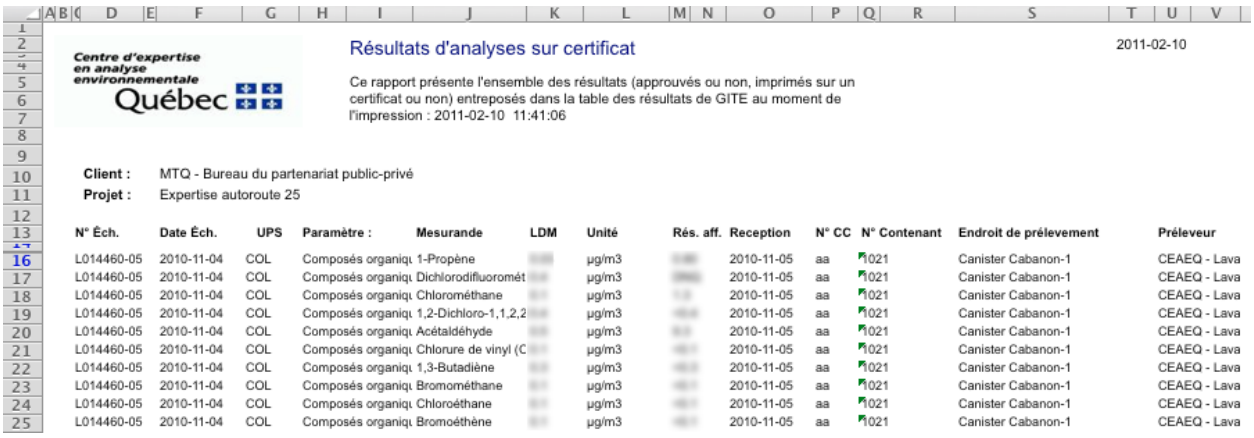
Figure 3.12: Couverture temporelle des données fournies par le MTQ pour les particules fines

Figure 3.13: Couverture temporelle des données fournies par le MTQ pour les particules en suspension

Composés organiques volatils

Les données COV du MTQ sont fournies pour deux stations, soit les stations C1 et C2 (voir Figure 3.3 pour la position de celles-ci), et se présentent sous la forme de tableaux au format Excel représentant l'ensemble des lectures pour toutes les stations sur une période donnée, typiquement de trois mois.

Pour chacune des lignes représentant une lecture unique, un ensemble d'informations est fourni, notamment la date du prélèvement dans la colonne "Date Éch.", la mesure prise dans la colonne "Rés. aff.", le paramètre analysé dans la colonne "Mesurande", la limite de détection pour celui-ci et son unité dans les colonnes "LDM" et "Unité". La colonne "Endroit de prélèvement" fournit finalement le nom de la station de mesure, le nom "Canister Cabanon-1" représentant la station C1, alors que "Canister Cabanon-2" est associé à la station C2. La Figure 3.14 présente l'ensemble de ces colonnes et un extrait des données. En tout, un total de 8418 lectures sont fournies pour ces deux stations entre les mois de mai 2010 et mai 2011 pour un ensemble de 78 COV.



Centre d'expertise en analyse environnementale Québec

Résultats d'analyses sur certificat

2011-02-10

Ce rapport présente l'ensemble des résultats (approuvés ou non, imprimés sur un certificat ou non) entreposés dans la table des résultats de GITE au moment de l'impression : 2011-02-10 11:41:06

Client : MTQ - Bureau du partenariat public-privé
Projet : Expertise autoroute 25

N° Éch.	Date Éch.	UPS	Paramètre :	Mesurande	LDM	Unité	Rés. aff.	Reception	N° CC	N° Contenant	Endroit de prélevement	Préleveur
L014460-05	2010-11-04	COL	Composés organiqu	1-Propène		µg/m3		2010-11-05	aa	1021	Canister Cabanon-1	CEAEQ - Lava
L014460-05	2010-11-04	COL	Composés organiqu	Dichlorodifluorométh		µg/m3		2010-11-05	aa	1021	Canister Cabanon-1	CEAEQ - Lava
L014460-05	2010-11-04	COL	Composés organiqu	Chlorométhane		µg/m3		2010-11-05	aa	1021	Canister Cabanon-1	CEAEQ - Lava
L014460-05	2010-11-04	COL	Composés organiqu	1,2-Dichloro-1,1,2,2		µg/m3		2010-11-05	aa	1021	Canister Cabanon-1	CEAEQ - Lava
L014460-05	2010-11-04	COL	Composés organiqu	Acétaldéhyde		µg/m3		2010-11-05	aa	1021	Canister Cabanon-1	CEAEQ - Lava
L014460-05	2010-11-04	COL	Composés organiqu	Chlorure de vinyl (C		µg/m3		2010-11-05	aa	1021	Canister Cabanon-1	CEAEQ - Lava
L014460-05	2010-11-04	COL	Composés organiqu	1,3-Butadiène		µg/m3		2010-11-05	aa	1021	Canister Cabanon-1	CEAEQ - Lava
L014460-05	2010-11-04	COL	Composés organiqu	Bromométhane		µg/m3		2010-11-05	aa	1021	Canister Cabanon-1	CEAEQ - Lava
L014460-05	2010-11-04	COL	Composés organiqu	Chloroéthane		µg/m3		2010-11-05	aa	1021	Canister Cabanon-1	CEAEQ - Lava
L014460-05	2010-11-04	COL	Composés organiqu	Bromoéthène		µg/m3		2010-11-05	aa	1021	Canister Cabanon-1	CEAEQ - Lava

Figure 3.14: Échantillon des données de COV du MTQ pour la période de décembre 2010 à février 2011

Les données de COV issues des fichiers fournis par le MTQ présentent quelques problèmes de gestion. Les mesures présentées sont en effet fournies de façon brute sans traitement préalable, ce qui fait en sorte qu'un certain nombre de manipulations sont requises avant de procéder à l'utilisation des valeurs obtenues, notamment en raison de la présence de valeurs non numériques. La présence ces dernières complexifierait l'intégration avec les autres données qui sont toutes numériques, si bien que le choix a été fait de faire l'utilisation de conventions

généralement acceptées pour les associer à des valeurs numériques. [3] Un certain nombre de mesures sont fournies avec comme valeur DNQ (Détecté mais non quantifié) ou encore une valeur inférieure à la limite de détection (dénotée par le symbole "<"). Par convention, les notations DNQ peuvent être traitées comme étant égales à 1,5 fois la limite de détection, alors que les mesures inférieures à la limite de détection devraient être remplacées par une valeur de 0,5 fois cette limite de détection. Ce besoin d'effectuer des traitements pour un certain nombre de mesures fait en sorte que des vérifications doivent être faites sur chacune des entrées avant leur intégration dans le système d'information.

3.3.2 RSQA

Le RSQA possède un réseau de 18 stations sur l'île de Montréal effectuant en continu des mesures de qualité de l'air. Dans le cadre de ce projet, les données issues des trois stations les plus proches de l'axe de l'autoroute 25 sont utilisées afin de complémenter les données des six stations du MTQ. L'ensemble des stations du RSQA est présenté à la Figure 3.15, alors que la Figure 3.16 présente les trois stations pour lesquelles des données sont utilisées et les types de données qui y sont mesurés.

3. Ces normes ont été fournies lors de communications directes avec les responsables de l'administration de ces données au MTQ.

Figure 3.15: Stations du RSQA

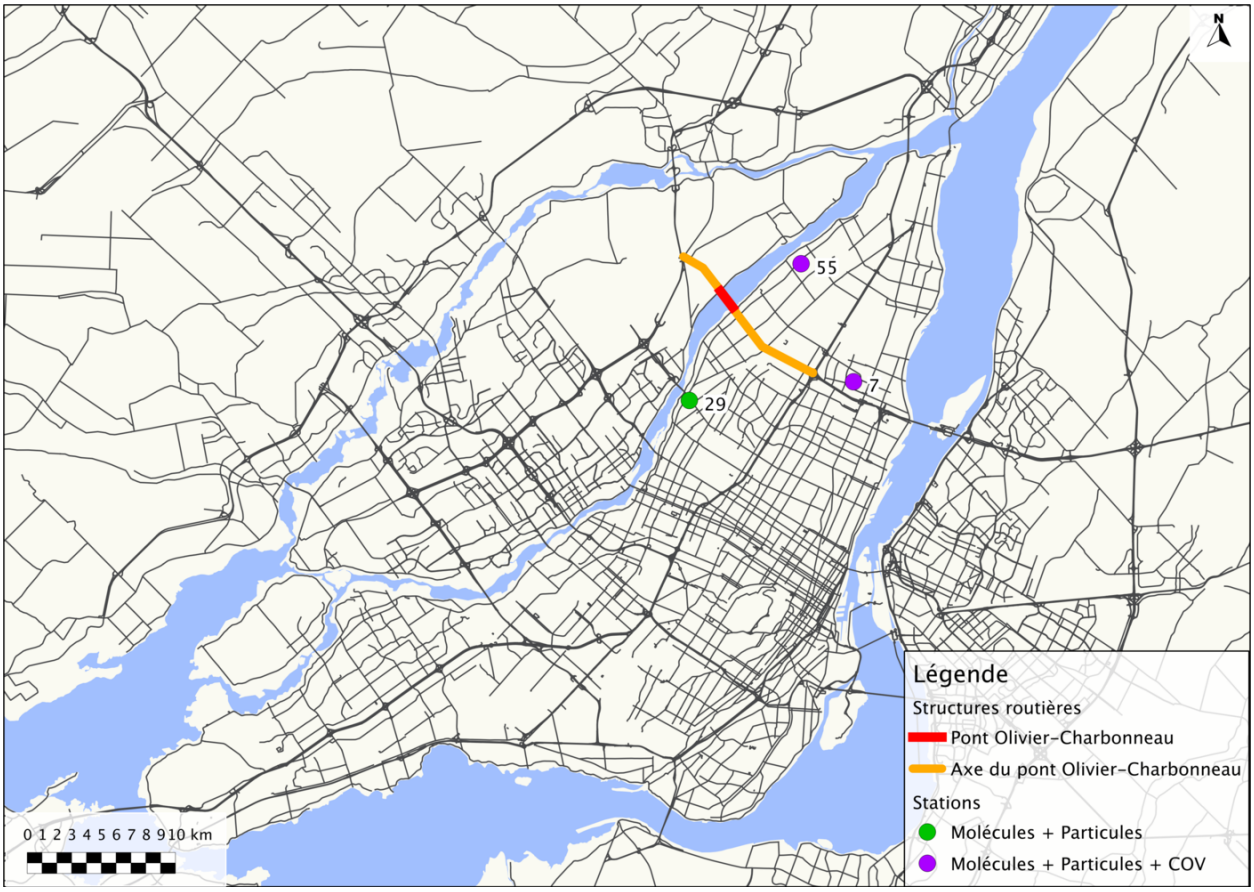


Figure 3.16: Stations utilisées du RSQA

Données de molécules et particules

Les données issues des stations du RSQA sont très similaires à celles obtenues pour les stations du MTQ. Les caractéristiques du point de vue des intervalles de collecte de données sont identiques, de même que les unités utilisées pour l'ensemble des paramètres. Les différentes molécules et le nombre de lectures prises pour chacune des stations sont présentés au Tableau 3.5.

La structure de la base de données fournie diffère toutefois grandement de celle détaillée précédemment et présente une structure mieux organisée. Le fichier fourni ne comprend en effet que deux tables, chacune représentant l'ensemble des lectures pour les trois stations disponibles pour l'année 2010 et l'année 2011. Les informations contenues dans chacune des tables sont par conséquent plus détaillées que dans le premier ensemble de données et incluent des informations quant à la station de provenance (colonne "POSTE") et au paramètre mesuré ("GAZ"). Les valeurs utilisées dans la colonne "GAZ" sont détaillées dans un fichier d'accom-

pagnement permettant de les lier à un paramètre. La définition des codes numériques issus de ce document et utilisés dans la base de données est présentée au Tableau 3.4. La Figure 3.17 présente quant à elle un échantillon de la structure des données fournies. Comme pour les données obtenues pour les stations du MTQ, des représentations graphiques de la couverture temporelle des données ont été produites et sont présentées aux Figures 3.19, 3.20, 3.21, 3.22, 3.23, 3.24, 3.25, 3.26 et 3.27, les lignes horizontales de chacun des graphiques représentant respectivement les stations 29, 55 et 7.

GAZ	POSTE	TEMPS	VALEUR	TYPE
1	7	2010-01-01		N
9	7	2010-01-01		N
10	7	2010-01-01		N
4	29	2010-01-01		N
8	29	2010-01-01		N
9	29	2010-01-01		N
10	29	2010-01-01		N
11	29	2010-01-01		N
8	55	2010-01-01		N
11	55	2010-01-01		N
1	7	-01-01 01:00:00		N
9	7	-01-01 01:00:00		N
10	7	-01-01 01:00:00		N
4	29	-01-01 01:00:00		N
8	29	-01-01 01:00:00		N
9	29	-01-01 01:00:00		N
10	29	-01-01 01:00:00		N
11	29	-01-01 01:00:00		N

Figure 3.17: Forme des données des fichiers du RSQA

Tableau 3.4: Codes numériques utilisés par le RSQA et les paramètres associés

Code numérique	Paramètre mesuré
1	SO_2
4	CO
8	O_3
9	NO_2
10	NO
11	PM_{25} (FDMS)
22	PM_{25} (GRIMM)
23	PM_{10} (GRIMM)
24	PST (GRIMM)

Les défis d'intégration de ces données sont moins complexes que pour les fichiers du MTQ puisque la structure présentée est constante et plus détaillée. Il est toutefois essentiel de faire appel à des dictionnaires externes pour connaître toutes les informations associées à une mesure, ce qui fait en sorte que les données ne peuvent pas être intégrées directement sans interventions pour chacune des entrées afin de les associer à leurs définitions.

Par ailleurs, plusieurs problèmes identifiés dans le cas des données du MTQ sont ici aussi présents et font en sorte que des manipulations doivent être faites sur chacune des entrées, notamment en ce qui concerne le changement d'heure, les unités variables en fonction des paramètres et l'absence d'information lorsqu'une mesure est jugée invalide sur une heure.

Un autre problème spécifique existe avec les données de la station 55 et le paramètre "PM10", les informations pour celui-ci étant présentées dans un fichier séparé. Les données de particules pour la station 55 sont en effet fournies dans un fichier au format Microsoft Excel avec des échantillonnages à des intervalles de trois jours pour l'année 2010, pour un total de 116 mesures. La feuille en question ne comprend que deux colonnes, soit "DATE" représentant la date de la lecture, ainsi qu'une colonne nommée "PM10" fournissant les lectures en $\mu g/m^3$. Le fichier lui-même comprend des données pour chacune des trois stations qui sont redondantes avec les informations fournies dans la base de données Access, ce qui fait en sorte qu'uniquement les mesures sur les particules doivent être extraites afin d'être intégrées dans le système d'information.

Composés organiques volatils

Les données de COV sont fournies pour deux stations du RSQA, soit les stations 7 et 55, dont la position peut être observée à la Figure 3.16. Les polluants mesurés sont dans la plupart des cas les mêmes que ceux obtenus des stations du MTQ, mais sont présentés dans des formats différents. Encore une fois ici, des feuilles de calcul du logiciel Microsoft Excel sont utilisées pour présenter les mesures faites par les stations. Un total de 21715 lectures sont fournies sur 164 polluants pour la station 7 et 204 polluants pour la station 55, représentant l'ensemble des mesures des années 2010 et 2011.

Les fichiers fournis sont toutefois beaucoup moins ordonnés que dans le cas des stations du MTQ. En effet, les données sont pour la plupart rendues disponibles dans des feuilles de tableurs comprenant aussi des informations redondantes déjà présentées dans les bases de données des molécules et particules. La nomenclature des fichiers est aussi problématique dans de nombreux cas et ne reflète pas toujours clairement le contenu. Par ailleurs, le format de ces feuilles de calcul est très variable en fonction des polluants étudiés. Les lectures des stations sont dans les faits séparés en trois, selon le type de COV étudié, soit des fichiers comprenant les données pour les COV polaires, les COV non polaires et les données Hazardous Air Polluants (HAP). L'échantillon présenté à la Figure 3.18 fait état du premier format dans lequel se présentent les données, avec la date de prise de mesure en en-tête et le paramètre étudié en colonne. Les fichiers stockant les mesures HAP sont plutôt structurés avec la date de la prise de mesure dans les colonnes et les polluants étudiés en en-tête. Cette différence entre les fichiers fait en sorte qu'un traitement manuel sera requis avant de procéder à toute forme d'opération automatisée sur les données.

	A	B	C	D	E
1	VOC Concentrations (ug/m3) at Montreal - 8200A Rue Chenier, Anjou NAPS No. 50133				
2	Compounds				
3	-	02-janv-10	08-janv-10	14-janv-10	20-janv-10
4	Ethane				
5	Ethylene				
6	Acetylene				
7	Propylene				
8	Propane				
9	1-Propyne				
10	Isobutane				
11	1-Butene/Isobutene				
12	1,3-Butadiene				
13	Butane				
14	trans-2-Butene				

Figure 3.18: Échantillon des données de COV pour l'année 2010 à la station 7 du RSQA.

Par ailleurs, les données fournies comprennent certains autres problèmes, notamment sous la forme de doublons. En effet, les données HAP pour la station 55 en 2010 sont fournies dans deux fichiers semblables, un de ces deux fichiers présentant des données sur une plus grande période. Les dates, dans ce cas précis, changent aussi de format à mi-chemin dans les lectures, ce qui fait en sorte qu'un autre traitement manuel devra être effectué afin de normaliser la structure de données. Les fichiers HAP présentent une autre problématique à gérer puisqu'ils incluent des colonnes de statistiques sur les mesures qui sont non documentées et ainsi inutilisables dans le cadre de ce projet. Ces colonnes supplémentaires devront par conséquent être éliminées avant tout traitement automatisé. Finalement, un des fichiers présente aussi des mesures indiquant des valeurs négatives ou inférieures à un seuil, mais contrairement aux stations du MTQ, aucune information n'a pu être obtenue définissant comment gérer ces cas. Il est important ici de noter que les deux fichiers portent la mention "Données incomplètes", mais qu'aucun fichier plus à jour n'a été rendu disponible.

Dans le cas d'autres fichiers, des valeurs de concentrations moyennes sont déjà calculées et devront être éliminées avant toute insertion dans le système d'information. Les fichiers non polaires présentent aussi la particularité de n'avoir aucune entrée pour certains polluants fournis, alors qu'aucune documentation ne permet d'expliquer les raisons de telles absences.

Plusieurs autres éléments méritent d'être mentionnés comme sources potentielles d'erreurs de traitement. Tout d'abord, certains fichiers définissent les unités de mesure dans les entêtes de colonnes, alors que d'autres définissent une unité pour l'ensemble du fichier dans des cellules supplémentaires avec des positions variables d'un fichier à l'autre, ce qui rend tout traitement automatisé impossible. Si la grande majorité des lectures sont faites avec $\mu g/m^3$ comme unité, les données HAP divergent et présentent des données en ng/m^3.

Les intervalles de temps entre les mesures sont quant à eux variables, avec des intervalles répertoriés de 6, 12 et même 18 jours. Cette variabilité des intervalles entre les lectures faites par les stations n'est pas documentée et ne peut par conséquent pas être expliquée.

Il est important de noter aussi des différences de nomenclatures et de niveau d'agrégation qui existent entre les données du MTQ et du RSQA. Par exemple, alors que les données du RSQA présentent des mesures pour le composé "m+p+o Xylènes", les stations du MTQ présentent les données sous les formes plus simples "m+p Xylènes" et "o Xylènes". La gestion de telles situation demande une connaissance approfondie des polluants étudiés, ce qui fait en sorte que cette étape devrait préférablement être effectuée lors de l'analyse des données plutôt que leur insertion dans le système d'information. Pour l'exemple mentionné ci-haut, dans le cadre de la production de tableaux statistiques, il serait donc nécessaire de procéder à l'addition des mesures des composés "m+p Xylènes" et "o Xylènes" pour obtenir la mesure de "m+p+o Xylènes" puisque sans cette intervention, l'absence de mesures agrégées pourrait

donner l'impression que ce polluant n'est pas étudié par la station en question.

Un dernier problème se présente sous la forme d'une nomenclature des composés différente de celle utilisée pour les stations du MTQ. En effet, les stations du RSQA produisent des résultats utilisant la nomenclature anglophone des polluants étudiés plutôt que la nomenclature francophone. Par conséquent il est essentiel de procéder à une conversion de ces polluants vers une nomenclature unifiée, soit la nomenclature francophone.

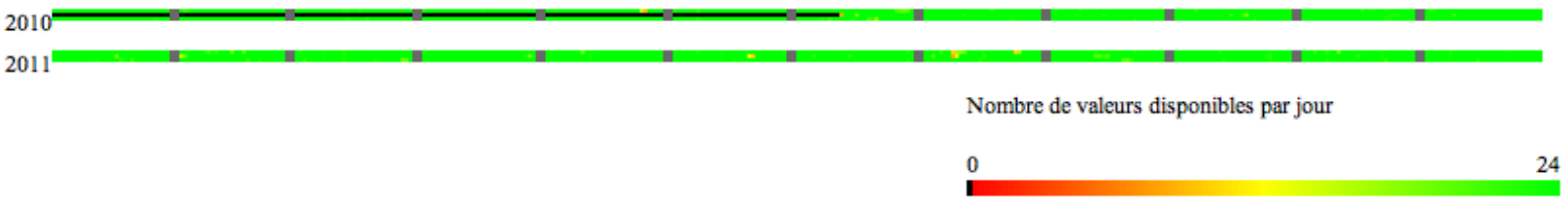


Figure 3.19: Couverture temporelle des données fournies par le RSQA pour le dioxyde d'azote

Figure 3.20: Couverture temporelle des données fournies par le RSQA pour le monoxyde d'azote

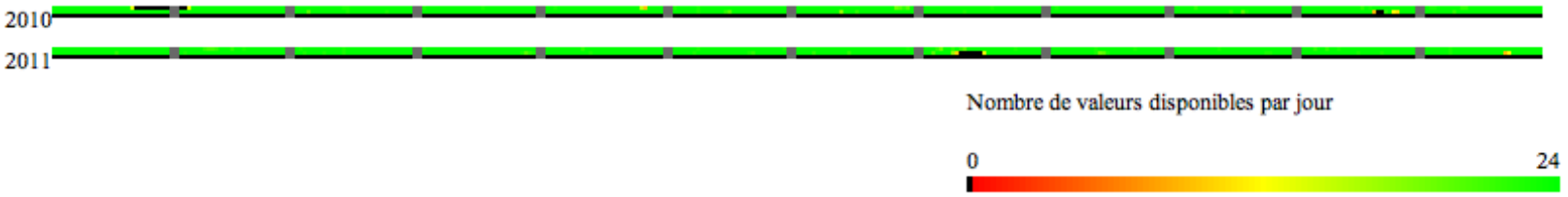


Figure 3.21: Couverture temporelle des données fournies par le RSQA pour l'ozone

Figure 3.22: Couverture temporelle des données fournies par le RSQA pour le monoxyde de carbone

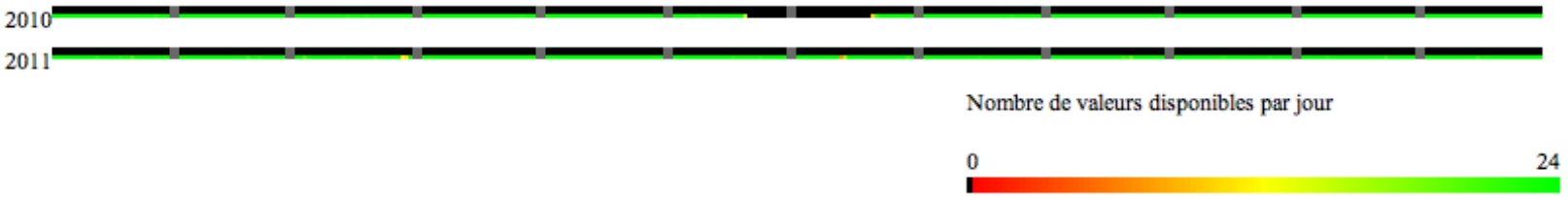


Figure 3.23: Couverture temporelle des données fournies par le RSQA pour le dioxyde de soufre

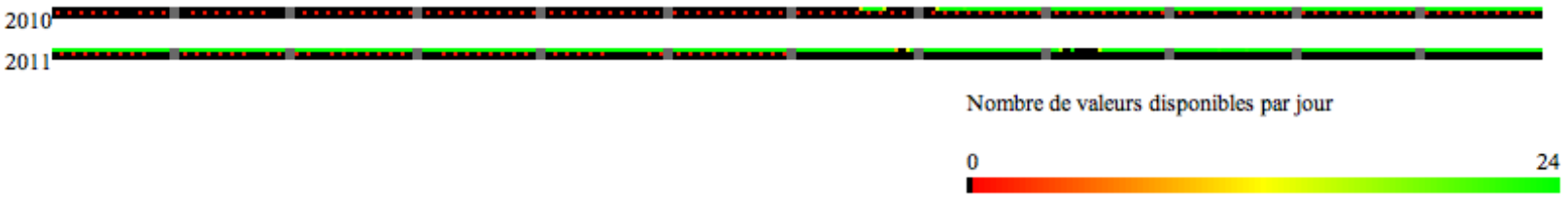


Figure 3.24: Couverture temporelle des données fournies par le RSQA pour les particules

Figure 3.25: Couverture temporelle des données fournies par le RSQA pour les particules fines (méthode GRIMM)

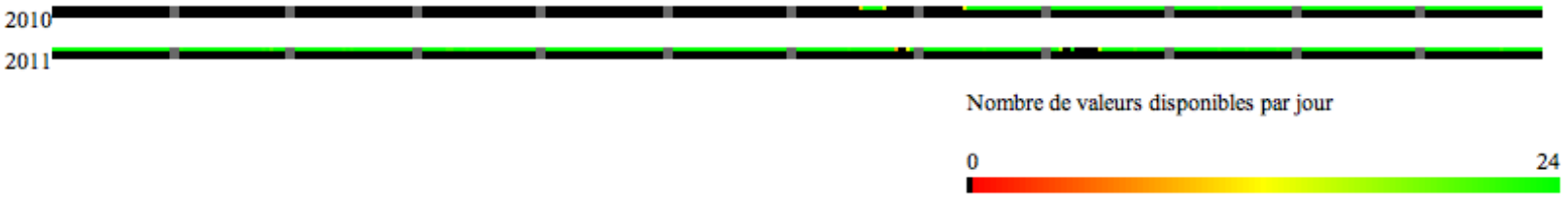


Figure 3.26: Couverture temporelle des données fournies par le RSQA pour les particules en suspension

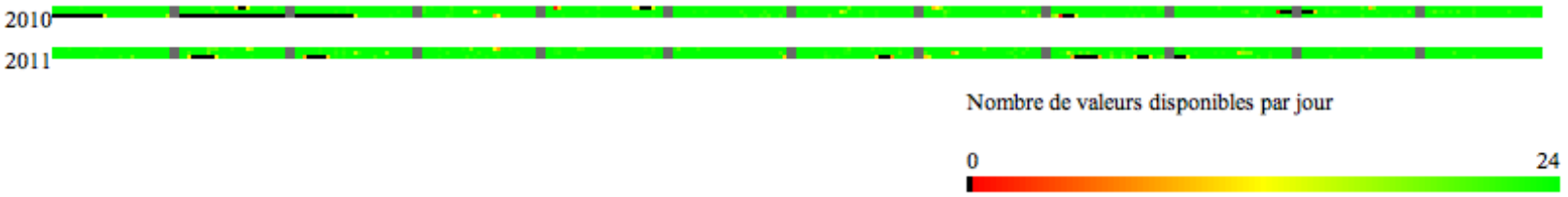


Figure 3.27: Couverture temporelle des données fournies par le RSQA pour les particules fines (méthode FDMS)

Tableau 3.5: Nombre de données disponibles par station et par paramètre étudié pour les stations de qualité de l'air du RSQA

Stations	NO2	NO	NOx	O3	PM10 (GRIMM)	PM25 (GRIMM)	PST (GRIMM)	CO	SO2	PM25 (FDMS)
29 - Parc Pilon - Montréal-Nord	17 354	17 354	0	17 104	12 203	12 267	12 016	17 412	0	17 088
55 - Rivière-des-prairies	12 742	12 742	0	17 152	168	0	0	9 560	0	17 057
7 - Chénier, Anjou	17 448	17 448	0	0	0	0	0	0	16 685	15 014

3.4 Données météorologiques

Les données météorologiques obtenues sont issues d'une seule station et sont fournies par le MDDEP dans un format unique, sous la forme d'une lecture par heure pour un ensemble de paramètres. La position de cette station est présentée à la Figure 3.28.

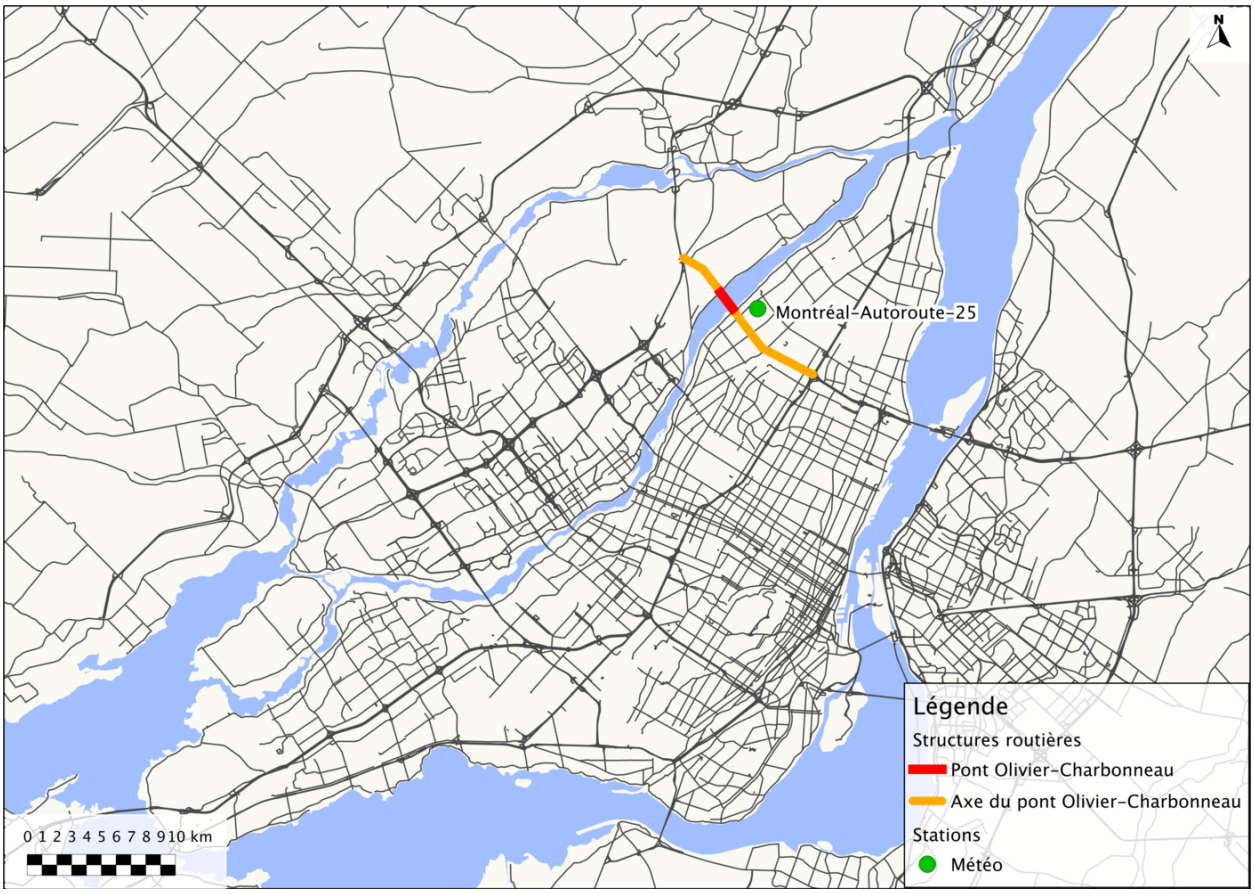


Figure 3.28: Station météo fournie

Le format se résume à un ensemble de fichiers texte sous forme de colonnes, chacune des colonnes utilisant un nombre prédéfini de caractères. Chaque ligne contient un numéro de station, suivi d'une date, suivi d'un code permettant d'identifier un paramètre et finalement de 24 colonnes représentant les informations pour les 24 heures de cette journée. Une nouvelle ligne signifie la fin des données pour cette journée et pour ce paramètre. Un fichier d'accompagnement permet d'obtenir les codes numériques pour identifier les paramètres, ainsi que les unités utilisées pour chacun de ceux-ci. Le dictionnaire relatant le contenu de chacune des colonnes est présenté au Tableau 3.6. Les informations complémentaires fournies permettent aussi de déterminer que les heures utilisées représentent la fin des périodes de mesure, et que

l'heure civile est utilisée pour l'ensemble des lectures.

Tableau 3.6: Dictionnaire permettant d'interpréter les fichiers de données météo

Position début	Position fin	Contenu
1	7	Numéro de la station
9	12	Année
14	15	Mois
17	18	Jour
20	24	Numéro de définition de donnée
26	26	Statut d'approbation
28	33	Données à 00h HNE
35	35	Statut de la donnée de 00h HNE
...	...	6 Caractères/1 Caractère pour les mesures et leur statut entre 01h et 22h
235	240	Données à 23h HNE
242	242	Statut de la donnée de 23h HNE

En tout, un ensemble de 83575 mesures de mars 2010 à mai 2011 a été obtenu dans un ensemble de 12 fichiers, l'envoi de ceux-ci ayant été fait en deux parties. La Figure 3.29 présente un échantillon du format de données, alors que le Tableau 3.7 fait état des codes et unités des différents paramètres.

```
7025237 2010 03 22 01080 1
                                0007.7 5 0007.2 1 0006.6 1 0006.5 1 0006.2 1 0005.7 1 0005.2 1 0005.5 1<CR>
7025237 2010 03 23 01080 1 0005.6 1 0005.0 1 0004.2 1 0003.8 1 0003.3 1 0002.7 1 0001.2 1 0000.5 1 0000.5 1 -000.7 1 -000.3 1 -000.3
1 -000.6 1 0000.2 1 -000.2 1 0000.8 1 0000.5 1 0000.9 1 0000.1 1 0000.4 1 -000.1 1 0000.2 1 0000.4 1 0000.5 1<CR>
7025237 2010 03 24 01080 1 -000.4 1 -000.3 1 -000.7 1 -000.8 1 -000.3 1 -000.3 1 -000.8 1 0000.2 1 0001.0 1 0002.1 1 0003.3 1 0005.8
1 0007.2 1 0009.0 1 0010.1 1 0011.4 1 0011.4 1 0011.1 1 0009.7 1 0008.3 1 0006.4 1 0004.6 1 0003.3 1 0003.8 1<CR>
7025237 2010 03 25 01080 1 0003.2 1 0003.7 1 0002.5 1 0002.0 1 0000.6 1 0001.3 1 0001.5 1 0003.1 1 0004.5 1 0006.8 1 0009.0 1 0011.5
1 0012.7 1 0013.8 1 0013.5 1 0013.3 1 0012.6 1 0012.2 1 0007.7 1 0006.5 1 0003.7 1 0001.4 1 -000.6 1 -002.0 1<CR>
7025237 2010 03 26 01080 1 -003.2 1 -004.8 1 -006.2 1 -007.2 1 -008.6 1 -009.6 1 -010.1 1 -010.0 1 -009.9 1 -008.6 1 -007.8 1 -006.4
1 -004.8 1 -003.2 1 -003.5 1 -001.8 1 -001.5 1 -001.5 1 -002.4 1 -003.3 1 -003.8 1 -004.6 1 -005.3 1 -005.8 1<CR>
7025237 2010 03 27 01080 1 -006.3 1 -006.4 1 -007.6 1 -008.1 1 -008.5 1 -009.0 1 -008.8 1 -008.1 1 -006.9 1 -005.3 1 -003.3 1 -002.1
1 -000.5 1 -000.9 1 0000.1 1 0001.2 1 0000.5 1 0001.0 1 0000.6 1 -000.3 1 -000.7 1 -000.3 1 -000.5 1 -000.8 1<CR>
```

Figure 3.29: Forme des fichiers de données météo

Comme pour les ensembles de données précédents, un certain nombre de problèmes concernent les données météorologiques fournies. Tout d'abord, toute automatisation devra, d'une façon semblable aux données de qualité de l'air, gérer la problématique des changements d'heure, l'ensemble des données étant fournies à l'heure normale de l'Est (UTC-5). De

plus, les données indiquent la fin d'une période de mesure, mais puisque le premier champ utilise une heure 0, la première mesure présentée pour une journée représente dans les faits la mesure entre 23:00:00 et 24:00:00 la journée précédente. Cette notation particulière devra être prise en considération lors de l'intégration des données au système d'information.

Les données étant fournies sous la forme de fichiers texte, la question de l'encodage des fichiers pourrait se révéler problématique lors de traitements automatisés. En effet, les fichiers fournis utilisent l'encodage US-ASCII, alors que la plupart des systèmes modernes, langages de programmation et SGBDR utilisent l'encodage UTF-8. Il pourrait donc être nécessaire de spécifier cet encodage lors de l'ouverture des fichiers, ou plus simplement de procéder à une conversion avant leur utilisation afin d'éviter des informations erronées sous la forme de caractères incorrects.

Malgré l'imposition d'une norme formelle décrite dans les fichiers d'accompagnement, il s'avère que celle-ci est incomplète et peut même parfois induire des erreurs. Par exemple, bien que le nombre de caractères pouvant être contenu sur une ligne soit défini par la norme, il s'avère que certaines lignes ne respectent pas exactement ce critère, ce qui peut mener à des erreurs lors de traitements automatisés. Par ailleurs, malgré la description d'un grand nombre de paramètres ainsi que des codes numériques associés, six paramètres sont fournis sans toutefois être documentés, soit les paramètres ayant les codes numériques 1148, 1196, 1151, 57854, 57855, 57856 et 57857. Si la plupart de ces paramètres ne possèdent pas d'entrées autres que des espaces vides dans les fichiers, les paramètres 1196 et 1151 peuvent être associés manuellement aux valeurs de vent maximum, le nom du fichier offrant une indication à cet effet, alors que des versions imprimables en fichier PDF permettent de conclure que le paramètre 1196 représente les valeurs de vitesse du vent et le paramètre 1151 la direction en degrés.

Une autre problématique n'engendrant toutefois pas de problèmes importants de gestion de données se présente sous la forme de plusieurs paramètres pour lesquels aucune valeur de mesure n'est fournie. Ces paramètres sont 1148, 11903, 1151, 57854, 57855, 57856 et 57857. En l'absence d'informations supplémentaires, il est impossible d'expliquer les raisons pour lesquelles ces paramètres sont fournis sans aucune mesure associée.

Finalement, une autre problématique qui ne pose pas de problèmes de gestion à ce point se présente sous la forme de plusieurs paramètres mesurant des données semblables fournies dans les mêmes fichiers. Par exemple, le fichier contenant les vitesses instantanées du vent présente les paramètres 1148, 1193, 1566 et 1192. Si les deux premiers paramètres ne comportent pas de mesures dans cet ensemble particulier, la possibilité d'avoir deux mesures pour la même information engendre des risques potentiels de création de doublons. Par conséquent, les paramètres ne possédant pas de mesures formelles devraient être ignorés lors du traitement

afin d'éviter le stockage d'informations discordantes et tout traitement automatisé devrait être fait en anticipant ce problème éventuel.

Tableau 3.7: Codes numériques et définitions des données météo

code [4]	paramètre	unité
1076, 1077, 1699	Température maximale	°C
1078, 1079, 1700	Température minimale	°C
1080, 1081, 1702	Température instantanée	°C
1082, 1083, 1701	Température moyenne	°C
1010, 1753, 1758, 1780, 1840, 2047	Pluie (pluviomètre à pesée)	mm
1746, 2155	Pluie (pluviomètre à augets)	mm
1011, 1754, 1759, 1781, 1842, 2057	Neige (pluviomètre à pesée)	cm
1012, 1593, 1763, 1782, 1838, 2037	Précipitations totales (pluviomètres à pesée)	mm
1709	Précipitations totales (pluviomètres à augets)	mm
1007, 1764, 1777, 1779, 1837, 2036	Précipitations totales cumulées (pluviomètre à pesée)	mm
1708	Précipitations totales cumulées (pluviomètre à augets)	mm
51377	Précipitations totales cumulées (pluviomètres à augets chauffant)	mm
1148, 1566	Vent - Direction	°
1192, 1193	Vent - Vitesse	km/h
1344, 1694	Humidité relative	%
1568, 2152	Point de rosée calculé à partir de l'humidité relative	°C
1151	Vent maximum - Direction	°
1196	Vent maximum - Vitesse	km/h

4. Les paramètres obtenus et documentés sont surlignés en vert, les paramètres documentés mais ne contenant que des entrées vides sont surlignés en gris, les paramètres non documentés mais avec des données sont surlignés en jaune, et les paramètres non documentés et pour lesquels uniquement des entrées vides sont obtenues sont surlignés en rouge.

3.5 Données de circulation

Les données de circulation contiennent principalement des comptages de véhicules sur différentes structures et ont été fournies sous la forme de deux ensembles distincts, soit 19 fichiers regroupant les mesures pour un sous-ensemble de ponts pour les années 2008, 2010 et 2011 ainsi qu'un autre ensemble de 106 fichiers comportant des comptages pour un ensemble de points de comptage pour certains des ponts en plus d'un très grand nombre de stations de comptage temporaires.

3.5.1 Stations de comptage permanentes

Les stations permanentes de comptage sont situées sur un ensemble de ponts situés dans l'axe ou à proximité de l'axe du nouveau pont de l'autoroute 25 et effectuent des enregistrements de façon automatisée et continue. Les données se présentent sous la forme de fichiers texte de grande taille avec des valeurs séparées par des virgules, chacun des fichiers représentant l'ensemble des comptages sur un pont pour une année entière. Comme pour les données de météo, un encodage spécifique est utilisé, soit l'encodage ISO-8859-15, qui doit être défini lors de l'ouverture des fichiers afin d'éviter de potentielles erreurs d'affichage pour certains caractères.

La structure générale des fichiers présente une mesure par ligne, chaque ligne présentant un horodatage (colonne "dah_debut") indiquant, contrairement aux autres ensembles de données rencontrés jusqu'ici, le début de l'intervalle de mesure et étant fourni à l'heure civile. Les périodes de mesures sont quant à elles variables, les comptages étant disponibles pour des intervalles de 15 minutes ou 60 minutes. Une autre colonne indique le numéro de la voie ("cod_voie") pour laquelle la mesure est fournie, ce numéro répondant à un standard défini qui permet de déterminer la direction du flot de circulation sur chaque voie, les numéros de 1 à 5 indiquant la direction nord ou ouest et les numéros de 6 à 10 indiquant la direction sud ou est en fonction de l'orientation de la route. La colonne "nombre" indique quant à elle le total de véhicules ayant circulé sur l'intervalle sur la voie en question. Dans certains cas, cette mesure s'accompagne d'une classe de véhicule basée sur la longueur indiquée dans la colonne classe, et d'une classe de vitesse associée indiquée dans la colonne vitesse. Par conséquent, il y a en général plus d'une mesure pour un horodatage en raison des voies multiples, des classes de véhicules et des vitesses auxquelles ceux-ci circulaient. La Figure 3.30 présente un échantillon de ces données de circulation, alors que la Figure 3.31 illustre la position des différentes stations permanentes obtenues. Finalement, le Tableau 3.8 fait état du nombre d'entrées obtenues pour chacune des stations pour chaque année.

```
1   dah_debut,cod_voie,nombre,classe,vitesse
2   01JAN2010:00:00:00,01, ,Indéterminé,Indéterminée
3   01JAN2010:00:00:00,01, ,0.0 - 21.9 pieds,"40,1 - 50 km/h"
4   01JAN2010:00:00:00,01, ,0.0 - 21.9 pieds,"70,1 - 80 km/h"
5   01JAN2010:00:00:00,01, ,0.0 - 21.9 pieds,"60,1 - 70 km/h"
6   01JAN2010:00:00:00,01, ,0.0 - 21.9 pieds,"50,1 - 60 km/h"
7   01JAN2010:00:00:00,02, ,Indéterminé,Indéterminée
8   01JAN2010:00:00:00,02, ,0.0 - 21.9 pieds,"50,1 - 60 km/h"
9   01JAN2010:00:00:00,02, ,0.0 - 21.9 pieds,"40,1 - 50 km/h"
10  01JAN2010:00:00:00,02, ,0.0 - 21.9 pieds,"60,1 - 70 km/h"
11  01JAN2010:00:00:00,02, ,0.0 - 21.9 pieds,"70,1 - 80 km/h"
12  01JAN2010:00:00:00,03, ,Indéterminé,Indéterminée
13  01JAN2010:00:00:00,06, ,0.0 - 21.9 pieds,"50,1 - 60 km/h"
14  01JAN2010:00:00:00,06, ,0.0 - 21.9 pieds,"60,1 - 70 km/h"
15  01JAN2010:00:00:00,06, ,Indéterminé,Indéterminée
16  01JAN2010:00:00:00,06, ,0.0 - 21.9 pieds,"80,1 - 90 km/h"
17  01JAN2010:00:00:00,07, ,Indéterminé,Indéterminée
18  01JAN2010:00:00:00,07, ,0.0 - 21.9 pieds,"50,1 - 60 km/h"
19  01JAN2010:00:00:00,07, ,0.0 - 21.9 pieds,"70,1 - 80 km/h"
20  01JAN2010:00:00:00,07, ,0.0 - 21.9 pieds,"60,1 - 70 km/h"
21  01JAN2010:00:15:00,01, ,0.0 - 21.9 pieds,"60,1 - 70 km/h"
22  01JAN2010:00:15:00,01, ,0.0 - 21.9 pieds,"50,1 - 60 km/h"
```

Figure 3.30: Forme des fichiers de données de circulation pour les ponts

Figure 3.31: Stations de comptage de circulation (ponts)

Tableau 3.8: Nombre de mesures disponibles par station par année pour les stations de comptage permanentes

Stations	2008	2009	2010	2011	Total
De Gaulle	44 778	-	46 458	20 475	111 711
Le Gardeur	-	-	808 638	830 273	1 638 911
Lepage	449 353	-	231 342	1 371 741	2 052 436
Papineau	1 661 531	-	1 491 563	1 712 805	4 865 899
Pie IX	52 005	-	641 086	1 434 412	2 127 503
Tunnel	70 182	-	89 535	268 144	427 861
Viau	51 147	-	16 884	-	68 031
Total	2 328 996	-	3 325 506	5 637 850	11 292 352

La taille imposante des fichiers fait en sorte que l'identification d'un grand nombre de problématiques associées aux données est un processus ardu. Néanmoins, deux problèmes principaux concernant la qualité des données ont pu être identifiés. Un certain nombre de doublons ont par exemple pu être observés. Ceux-ci se présentent sous la forme de comptages doubles pour un intervalle de mesure et un paramètre donné. Les valeurs dans ces cas précis sont différentes et la raison de leur apparition ne peut être expliquée. De telles situations devront être prises en compte dans le cadre des traitements automatisés des informations afin d'éviter d'éventuels conflits.

Une autre situation qui a été observée en un point unique pourrait toutefois se révéler plus compliquée à gérer. Dans ce cas précis, les données sont fournies à la fois pour un intervalle de 1 heure et pour les différents intervalles de 15 minutes. Un tel cas de chevauchement engendrera inévitablement des problèmes lors de l'intégration dans le système d'information et lors de la mise en place de processus permettant de ramener les données à des intervalles de temps semblables, risquant ainsi de surévaluer grandement le nombre réel de véhicules. Le problème n'ayant pu être identifié qu'à un seul endroit, il semble tout de même improbable qu'il puisse être présent à grande échelle et ainsi engendrer des problèmes majeurs de qualité de données.

D'autres problématiques plus mineures peuvent aussi être associées aux ensembles de données de circulation pour les ponts. Par exemple les classes de vitesses et de tailles de véhicules sont variables d'un fichier à l'autre et parfois même à l'intérieur d'un seul fichier. Les variations pour les classes de vitesses sont mineures (par exemple 40-49.9 km/h contre 41.1-50 km/h), si bien qu'elles pourront être gérées facilement lors de l'insertion des données dans le système d'information. Toutefois, les variations dans les classes de véhicules sont plus prononcées, rendant l'intégration de toutes les données sur des bases communes impossible.

Les horodatages au moment des changements d'heure de mars envoient aussi des informations contradictoires quant au fuseau horaire dans lequel les données sont fournies. À ces dates, des comptages sont disponibles pour l'heure allant de 2:00 à 3:00, ce qui est contre-intuitif, cette heure, du point de vue de l'heure civile, n'existant pas. Des vérifications ont toutefois permis de confirmer que ces mesures sont des artefacts restant après des manipulations et que l'ensemble des mesures est bel et bien à l'heure civile. Idéalement, ces artefacts devraient être éliminés lors de l'ajout des données au système d'information.

Finalement, les mesures possédant des intervalles de mesures variables, il est difficile d'identifier spécifiquement l'intervalle de mesure pour une valeur en particulier. Cette valeur d'intervalle étant parfois nécessaire pour réaliser des analyses, des procédures impliquant des vérifications sur les horodatages des mesures précédentes et suivantes devraient permettre d'identifier l'intervalle et d'ajouter ce renseignement au système d'information.

Comme dans le cas des données de qualité de l'air, des graphiques illustrent la répartition temporelle des données fournies et sont présentées à la Figure 3.32. Chacune des lignes du graphique présente l'information pour une station en particulier, dans le même ordre que celui utilisé au Tableau 3.8. Toutes les stations présentent des périodes pour lesquelles aucune mesure n'est disponible, si bien qu'en aucun temps l'ensemble des stations n'est fonctionnel à un moment précis. À l'image des graphiques produits pour la qualité de l'air, les discontinuités dans les mesures tendent à survenir sur des périodes prolongées de plusieurs jours d'affilée.

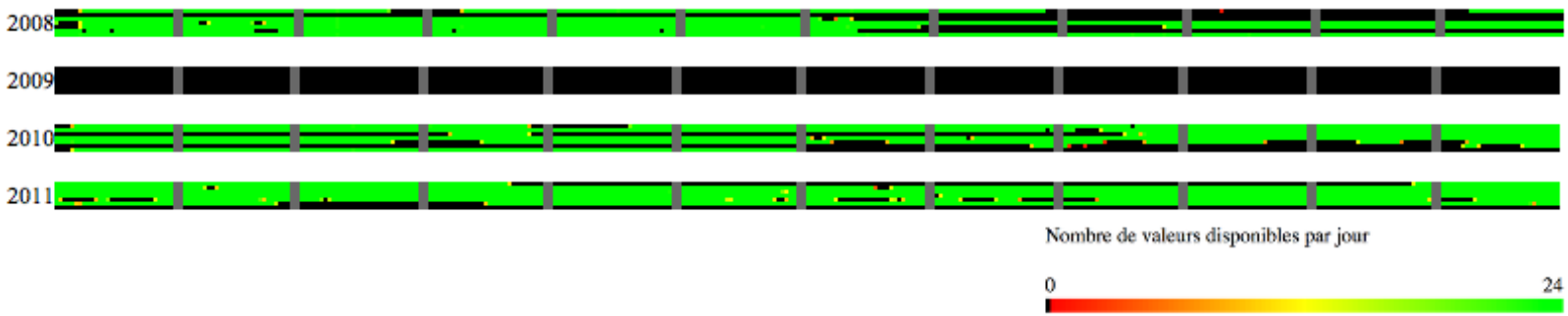


Figure 3.32: Couverture temporelle des stations de comptage permanentes

3.5.2 Autres sites de comptage

Un autre ensemble de données rassemblant des points de comptages dispersés sur un territoire à proximité de l'axe du nouveau pont a aussi été fourni, présentant des données de comptages de 2008 à 2011. Ces comptages comportent un ensemble disparate de fichiers comportant des mesures redondantes faites par certaines des stations pour les ponts mentionnés précédemment mais déjà agrégés à l'heure, quelques autres comptages automatisés et permanents ainsi qu'un ensemble de relevés manuels, notamment à certaines intersections importantes et sur des bretelles d'échangeurs autoroutiers. Il existe une superposition de ces comptages avec les mesures présentées à la Section 3.5.1. Celle-ci s'explique par le fait que les données détaillées ont été fournies plus tard que les données agrégées présentées ici. En raison de leur niveau de détails, les données désagrégées ont toutefois été préférées partout où c'était possible. Dans le cas du pont Le Gardeur, il est à noter que le fichier de données de 2008 n'est pas redondant avec les données précédemment étudiées pour ce pont et devra par conséquent être utilisé lors du peuplement du système d'information.

Un total de 104 fichiers comportant des mesures sont fournis, sous la forme de fichiers de tableur du logiciel Microsoft Excel, de fichiers textes, de documents du logiciel Microsoft Word ainsi que sous format PDF. La position géographique de l'ensemble de ces stations peut être observée à la Figure 3.33. Le nom de chacune des stations ainsi que le nombre d'entrées fournies et utilisées pour chacune d'entre elles, à l'exception des doublons, est présenté au Tableau 3.9. Contrairement aux autres ensembles de données présentant des défauts de continuité, il est difficile d'offrir une forme de visualisation acceptable de la couverture temporelle obtenue pour ces stations, notamment en raison de la courte période sur laquelle bon nombre d'entre elles recueillent les données.

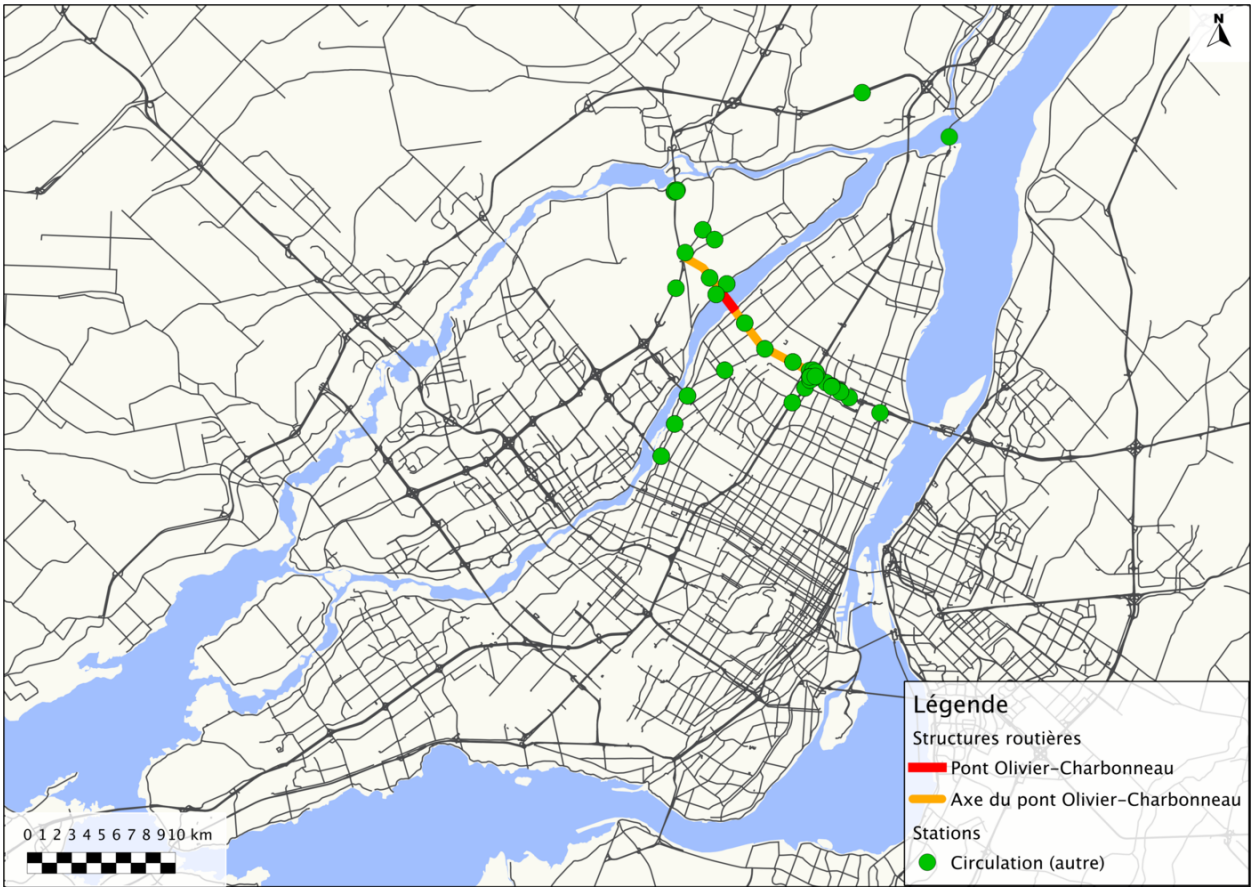


Figure 3.33: Positions des autres sites de comptage

Tableau 3.9: Nombre de données disponibles par station par année pour les stations de comptage temporaires

Station	Nombre d'entrées
Pont Le Gardeur	16 918
A-40 près de Galeries d'Anjou	14 120
A-40 Près de Langelier	2 322
A-25 au Nord de Montée Saint-François	938
A-640 Entre Montée Dumais et Montée des Pionniers	33 880
Montée Masson au Nord de Avenue Marcel Villeneuve	1 524
Bretelle Boulevard des Milles-Îles vers A-25S	909
Bretelle A-25S vers Boulevard des Milles-Îles	909
Suite à la page suivante	

Tableau 3.9: Nombre de données disponibles par station par année pour les stations de comptage temporaires (suite)

Station	Nombre d'entrées
Bretelle Boulevard des Milles-Îles vers A-25N	909
Bretelle A-25N vers Boulevard des Milles-Îles	911
Bretelle A-25S Sortie Beaubien	657
Bretelle A-25S Sortie Sherbrooke	2 209
Bretelle A-25S Sortie Souligny	1 821
Intersection Boulevard Henri-Bourassa Est et Boulevard Louis-H Lafontaine	3 504
Bretelle Montée Masson vers A-25N	144
Avenue Marcel Villeneuve (1km à l'est de Avenue Roger-Lortie)	1 516
Avenue Roger-Lortie (Intersection voie CP)	1 910
Boulevard Lévesque Est (300m à l'est de Avenue Roger-Lortie)	673
Boulevard Lévesque Est (400m à l'ouest de Avenue Roger-Lortie)	715
Boulevard Louis-H-Lafontaine au Nord de boulevard Maurice Duplessis	960
A-25 au Sud de A-40	423 592
A-25 au Sud de de Bombardier	11 950
Intersection Boulevard Henri-Bourassa Est et Boulevard Lacordaire	3 504
Intersection Boulevard Henri-Bourassa Est et Boulevard Saint-Michel	384
Intersection Boulevard Henri-Bourassa Est et Boulevard PieIX	1 122
Intersection Boulevard Henri-Bourassa Est et A-19	3 120
A-25N Échangeur Anjou	660
Bretelle A-40E vers A-25N	2 873
Bretelle A-40E vers A-25S	112
A-40E Échangeur Anjou	660
A-40E - Approche bretelle A-40E vers A-25N	1 140
Bretelle A-40O vers A-25N	280
Bretelle A-40O vers A-25S	112
Bretelle A-25N vers A-40E	112
Bretelle A-25N vers A-40O	352

Suite à la page suivante

Tableau 3.9: Nombre de données disponibles par station par année pour les stations de comptage temporaires (suite)

Station	Nombre d'entrées
Bretelle A-25S vers A-40O	2 168
Bretelle A-25S vers A-40E	3 248
A-40O Échangeur Anjou	1 140
Autoroute 25 au sud du Boulevard Wilfrid-Pelletier	2 016
Autoroute 25 au nord du Boulevard Yves-Prévost	2 016
Bretelle d'entrée Autoroute 25-N au sud du Boulevard Wilfrid-Pelletier	336
Bretelle de sortie Autoroute 25-N près du Boulevard Yves-Prévost	336
Total	548 862

Si le total des entrées à intégrer est de beaucoup inférieur aux ensembles de données concernant uniquement les ponts, les données ont malheureusement de très importants problèmes d'organisation qui rendent leur intégration difficile, voir impossible sans un traitement manuel de chacun des fichiers. En effet, pour les 104 fichiers, un peu plus de 20 formats de fichiers peuvent être identifiés, ce qui fait en sorte que le traitement automatisé nécessiterait l'adaptation à chacun de ces types de fichiers d'éventuels outils développés.

Le format le plus commun se présente sous la forme d'une feuille de calcul du logiciel Microsoft Excel, utilisée par environ près de 25 % des fichiers, et est présenté à la Figure 3.34. Chacune des lignes fait état d'un nombre de véhicules calculé (colonne "Débit") pour une voie ou une direction (colonne "Voie") à une date et une heure (colonnes "Date" et "Heure"). Un ensemble de sept autres fichiers utilise une structure très semblable, la colonne identifiant les valeurs mesurées étant remplacée par une colonne identifiée "INDÉTERMINÉ" et une colonne s'ajoutant indiquant le pourcentage de camions recensés sur ce nombre. Dans tous les cas, comme pour les données de circulation sur les ponts, les horodatages indiquent le début de l'intervalle sur laquelle la mesure s'applique, bien que certains fichiers spécifient clairement les heures de début et de fin, et ces horodatages utilisent le temps à l'heure civile.

	A	B	C	D	E	F	G
1	Section de trafic	Voie	SEMAINE	Jour	Date	Heure	Débit
2	'0002516000	'11	08-01-06	Lundi	08-01-07	12:00-12:59	
3	'0002516000	'11	08-01-06	Lundi	08-01-07	13:00-13:59	
4	'0002516000	'11	08-01-06	Lundi	08-01-07	14:00-14:59	
5	'0002516000	'11	08-01-06	Lundi	08-01-07	15:00-15:59	
6	'0002516000	'11	08-01-06	Lundi	08-01-07	16:00-16:59	
7	'0002516000	'11	08-01-06	Lundi	08-01-07	17:00-17:59	
8	'0002516000	'11	08-01-06	Lundi	08-01-07	18:00-18:59	
9	'0002516000	'11	08-01-06	Lundi	08-01-07	19:00-19:59	
10	'0002516000	'11	08-01-06	Lundi	08-01-07	20:00-20:59	
11	'0002516000	'11	08-01-06	Lundi	08-01-07	21:00-21:59	

Figure 3.34: Format de données le plus commun des stations de circulation

Parmi les problèmes potentiels identifiés, certains fichiers ne présentent qu'un code numérique sans informations supplémentaires, rendant difficile l'identification de la station. Dans d'autres cas, il est difficile d'identifier formellement le site du comptage, la seule information fournie étant une appellation générique. Certains autres fichiers ne présentent aucune date de collecte de données, ce qui rend l'utilisation des valeurs obtenues impossible.

Dans certains cas, c'est la nature même des fichiers fournis qui pose un défi de gestion. Par exemple, certaines structures de données se répètent dans plusieurs fichiers qui sont de types différents, ce qui fait en sorte que des formats qui seraient compatibles sont rendus incompatibles pour une raison inexplicable. Dans un cas précis, une même structure est présente dans un ensemble de fichiers qui compte des fichiers texte bruts, dans des documents du logiciel Microsoft Word ainsi que des documents de type PDF. D'autres variations mineures sont un problème récurrent pour les autres formats identifiés, par exemple des colonnes qui changent d'appellation ou les informations à propos du site de collecte de données qui changent de format d'un fichier à l'autre. Ces variations rendent très difficile tout traitement automatisé, ce qui fait en sorte que presque tous les fichiers devront faire l'objet de traitements manuels afin d'établir un standard fonctionnel.

Au-delà des problèmes structurels, d'autres problèmes déjà rencontrés pour les autres types de données sont aussi présents pour ces données de comptage. Notamment, l'accès à des définitions extérieures est bien souvent nécessaire afin de localiser ou nommer les stations. La présence de valeur de voies ou de plusieurs stations différentes à l'intérieur d'un même fichier rend cette tâche d'identification encore plus ardue.

Certains fichiers présentent aussi quelques cas de doublons internes, un cas spécifique ayant été identifié de façon spécifique pour la bretelle de sortie de l'Autoroute 25 Nord près du Boulevard Yves-Prévost. Finalement, dans le cas des fichiers textes, comme pour les ensembles précédents, des problèmes liés à l'encodage des fichiers peuvent subvenir, les

encodages ISO-8859-15 et US-ASCII étant utilisés.

3.6 Synthèse

L'analyse descriptive de l'ensemble des fichiers de données permet d'illustrer la taille imposante et la complexité inhérente à la gestion de très grands ensembles de données spatio-temporelles. Le Tableau 3.10 présente la synthèse des sections précédentes quant au nombre de fichiers et de formats présents dans l'ensemble des données reçues.

En plus de certains problèmes spécifiques à chaque format documenté, un ensemble de problématiques récurrentes pour presque tous les fichiers devront être prises en compte avant de procéder à une quelconque utilisation des données de différents types. Notamment, les questions de changements d'heures concernant les données de qualité de l'air et de météo font en sorte qu'un ajustement sur les horodatages devra avoir lieu sur de très grands ensembles de données. L'utilisation du début de l'intervalle de mesure comme horodatage pour les données de circulation alors que les autres types utilisent la fin de cet intervalle est aussi un problème auquel il faudra remédier afin d'obtenir une normalisation permettant de traiter l'ensemble des mesures de la même façon.

Tableau 3.10: Le nombre de fichiers et de formats de données pour chaque type

Type	Nombre de fichiers	Nombre de formats	Taille (Mo)
Qualité de l'air (MTQ)	1 Base Access (46 tables)	1	11,5
Qualité de l'air (RSQA)	1 Base Access (2 tables) / 1 Tableur Excel	2	19
COV (MTQ)	6 Tableurs Excel	1	3,4
COV (RSQA)	7 Feuilles Excel dans 6 fichiers (1 feuille redondante)	2	7
Météo	12 fichiers texte	1	1,3
Circulation (Ponts)	19 fichiers texte	1	668,6
Circulation (Autres)	104 fichiers (Excel, Word, texte, PDF)	21	62,7
Total	150 fichiers (207 sous-ensembles)	29	773,5

La taille des ensembles de données et les besoins de faire des traitements préalables à leur utilisation et leur insertion font en sorte qu'un certain nombre de procédures automatisées doivent être développées. Or, plusieurs fichiers ne sont pas adaptés à de tels traitements automatisés, et devront par conséquent être normalisés manuellement avant de pouvoir faire l'objet de telles procédures. Les problèmes liés à des informations entrées en double dans les fichiers, ainsi que l'encodage de ces mêmes fichiers, devront aussi être pris en considération avant de faire l'intégration des données.

La désorganisation dans plusieurs ensembles de données est une des principales motivations menant à la création d'un système d'information unifié. Suite à l'inventaire des ensembles de données et de leurs problématiques spécifiques, une schématisation de la forme que devrait prendre le système d'information peut être faite et est présentée à la Figure 3.35. Cette dernière reprend les couleurs utilisées précédemment, le bleu étant associé au phénomène du chaos de l'information, le jaune à la phase d'organisation et le vert à la phase d'exploitation des données. Le prochain chapitre traitera du développement, du montage et du peuplement du système d'information qui devrait, à terme, pallier les faiblesses identifiées et ainsi soutenir efficacement les analyses.

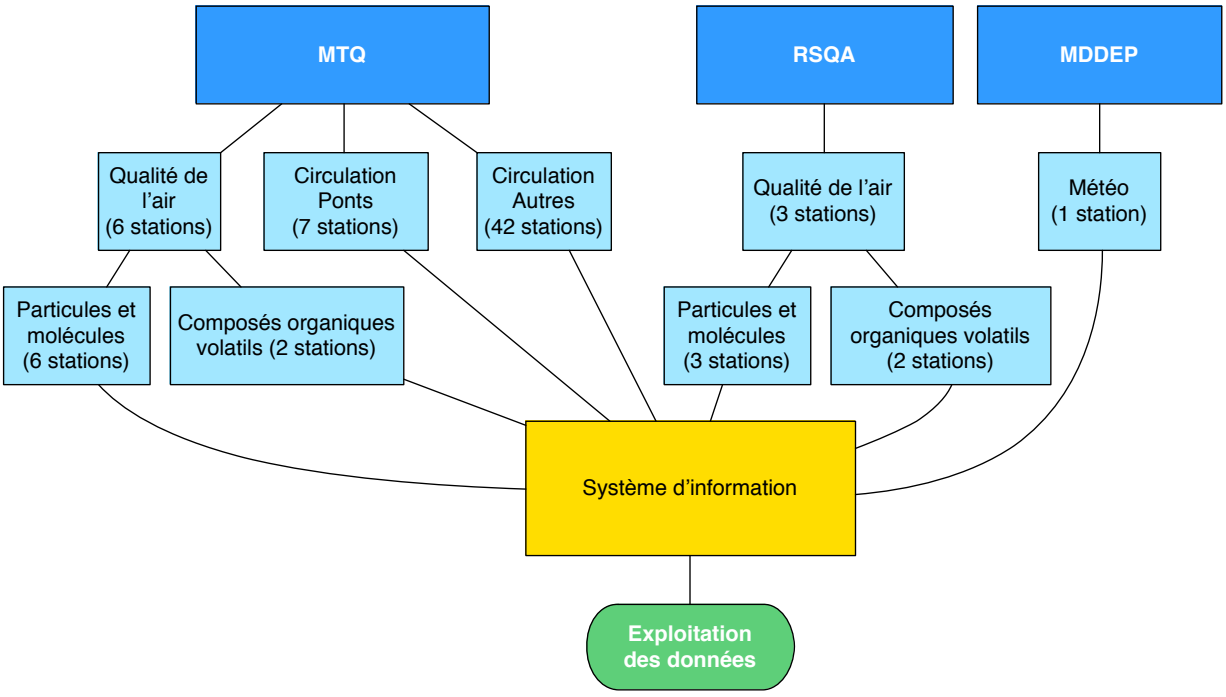


Figure 3.35: Structure du système d'information suite à l'inventaire des données disponibles

CHAPITRE 4

DÉVELOPPEMENT DU SYSTÈME D'INFORMATION

La taille des ensembles de données ainsi que leur complexité inhérente font en sorte que le développement d'un système d'information avancé est nécessaire afin d'offrir un soutien efficace aux analyses. Pour ce faire, différentes formes de systèmes d'informations doivent d'abord être étudiées afin d'éventuellement pouvoir développer la solution la plus appropriée et la plus efficace possible. Une telle solution, dans une approche orientée-objet, passe d'abord par la décomposition des différents ensembles de données en objets simples. Par la suite, un schéma de base de données permettant de stocker l'ensemble des informations pour chacun de ces objets doit être défini, ainsi qu'un ensemble de contraintes assurant la qualité des données stockées et de l'élaboration de standards stricts permettant de normaliser l'ensemble des informations stockées. Finalement, un ensemble de choix technologique doit être fait avant de procéder à l'implantation de la solution qui sera définie.

4.1 Options de stockage

L'application de la méthode de développement Agile implique de procéder au développement du système de façon itérative. Il est toutefois possible de cerner différentes options en ce qui a trait à la forme générale que devrait prendre le système d'information à développer. Les deux formes les plus communes sont celles de tables uniques comportant l'ensemble des informations, ou l'utilisation d'un système présentant l'information sous la forme d'objets simples et décomposant ainsi l'ensemble des données.

4.1.1 Système à table unique

Tel que présenté à la section précédente, chacun des ensembles de données appartenant à un type spécifique présente un ensemble de points communs les rendant plus ou moins compatibles malgré les différences existant dans la codification de l'information. Un traitement individuel pour chacun de ces types se révélerait par conséquent plus simple à gérer, les schémas pour chacun d'entre eux se limitant à un ensemble d'information limité. Par exemple, un schéma permettant d'emmagasiner toutes les informations ayant trait aux données de qualité de l'air pour les molécules et les particules pourrait se limiter à la structure présentée au Tableau 4.1.

Tableau 4.1: Schéma simple pour les données de qualité de l'air

Champ	Information entreposée
station	Code numérique de la station
paramètre	Le nom du paramètre
horodatage	Le moment de la prise de mesure
mesure	La mesure rapportée par la station

Un tel format permettrait d'emmagasiner, à l'image de la structure des données fournies par le RSQA, l'ensemble des informations contenues à l'intérieur des fichiers de qualité de l'air à l'exception des données de dictionnaires accompagnant les fichiers, qui incluent notamment la position des stations et les unités de mesure. Puisque le nombre de stations et de paramètres est faible, le stockage du nom complet ou de la position de chacune des stations peut être effectué de façon externe sans problèmes de complexité importants. Le nombre total de mesures de qualité de l'air étant lui aussi relativement faible, l'ajout de ces informations à chaque entrée engendre une redondance qui peut tout de même être gérée assez simplement puisque les informations sont limitées à un type précis de données clairement défini et au format constant.

Si l'ensemble des informations devait être stocké, des problèmes pourraient commencer à apparaître et présenter une redondance inutile à chaque enregistrement, sans toutefois créer de problèmes majeurs de gestion. Une telle solution prendrait essentiellement la forme rapportée au Tableau 4.2.

Tableau 4.2: Un schéma contenant l'ensemble des informations des données de qualité de l'air

Champ	Information entreposée
station	Code numérique de la station
source	Propriétaire de la station
fichier_source	Le fichier duquel est extrait la mesure
paramètre	Le nom du paramètre
horodatage	Le moment de la prise de mesure
mesure	La mesure rapportée par la station
latitude	La latitude de la station
longitude	La longitude de la station
unité	L'unité de la mesure

L'accès aux données dans une telle structure se fait assez aisément, mais certaines opérations se révéleraient plus ardues. Si, par exemple, une station avait été mal codifiée, il serait en effet nécessaire de procéder au changement des informations relatives à cette station sur plusieurs dizaines ou centaines de milliers d'enregistrements. L'augmentation du nombre de stations et de paramètres aurait aussi pour effet de complexifier graduellement les opérations de gestion de l'ensemble des données. Si la gestion de 9 stations et de 10 paramètres est plutôt simple, l'insertion des données de COV et des données pour l'ensemble de toutes les stations du RSQA ferait passer le nombre de stations et de paramètres à 24 et 243 respectivement.

La multiplication des types de données et d'informations spécifiques à chacun de ces types a aussi pour effet de complexifier la structure d'une table unique et de rendre plus difficile la gestion des conflits sous la forme de doublons énoncés dans le chapitre précédent. Par exemple, l'ajout de l'ensemble des données de circulation engendrerait l'insertion de plusieurs dizaines de stations, en plus de forcer une augmentation du nombre de champs pour permettre le stockage de la voie, de la direction, de la classe de véhicule et de la classe de vitesse. En outre, les intervalles de mesure variables pour ce type de données ajouteraient un champ supplémentaire pour l'ensemble des données. La représentation d'un schéma de table unique permettant de stocker toutes ces informations prendrait alors la forme du Tableau 4.3.

Tableau 4.3: Un schéma contenant l'ensemble des champs nécessaires pour stocker l'ensemble des informations

Champ	Information entreposée
station	Code numérique de la station
source	Propriétaire de la station
fichier_source	Le fichier duquel est extrait la mesure
paramètre	Le nom du paramètre
horodatage	Le moment de la prise de mesure
mesure	La mesure rapportée par la station
latitude	La latitude de la station
longitude	La longitude de la station
unité	L'unité de la mesure
voie	Le numéro de la voie de circulation
direction	La direction dans laquelle un comptage est effectué
vitesse	La classe de vitesse des véhicules comptés
classe_vehicules	La classe de véhicules comptés
intervalle	L'intervalle sur lequel la mesure s'applique

L'ajout de champs qui stockent des informations complètes à chaque entrée, pour des millions d'entrées, engendre des problèmes de redondance importants, qui peuvent se révéler coûteux en terme de performance pour le système. Par ailleurs, une telle structure rend difficile de faire un traitement égal de toutes les données, les données de qualité de l'air devant être obtenues en faisant une requête sur le champ "parametre" alors qu'une requête sur des données de circulation porterait plutôt sur les champs "vitesse" et "classe_véhicules". Elle serait toutefois essentielle en regard du besoin de stocker plusieurs paramètres pour chacun des enregistrements de circulation. De telles divergences dans les méthodes d'accès aux types de données vont à l'encontre des objectifs de normalisation énoncés.

L'utilisation de plusieurs champs pour stocker des informations de natures similaires vient aussi complexifier la tâche de développement d'un système de contrainte qui assure la cohérence des données stockées. Une telle contrainte viserait, dans ce cas précis, les champs "station", "paramètre", "horodatage", "voie", "direction", "vitesse" et "classe_vehicules" afin d'assurer qu'un seul enregistrement ne puisse être stocké pour une station et un paramètre précis.

La structure en question engendre aussi de la confusion pour les éventuels utilisateurs et peut ainsi entraîner des erreurs humaines. Par exemple, il serait possible pour un utilisateur d'hésiter à placer une information sur la vitesse du vent alors qu'un champ vitesse existe.

En résumé, une telle structure, en plus d'engendrer des quantités importantes d'informations redondantes, rend les opérations de maintenance plus difficile tout en ajoutant des éléments pouvant générer de la confusion pour les utilisateurs qui en font l'usage en raison de l'appartenance de certains champs à des types de données spécifiques.

Il est aussi possible de définir une table indépendante pour chaque type de données disponibles. Ainsi, la multiplication des champs peut être contenue si seulement ceux qui sont appropriés sont conservés. Une telle solution présente toutefois elle aussi des problèmes d'organisation. Tout d'abord, le nombre de points d'accès se voit augmenté au fur et à mesure que de nouveaux types de données sont ajoutés. La séparation de l'information s'accompagne aussi de difficulté d'accès à l'information, une nouvelle requête devant être écrite pour chaque nouveau type de données, même si la requête est de même nature. Par ailleurs, des problèmes mentionnés plus tôt, comme la redondance de certaines informations et le temps requis pour modifier des valeurs associées à de nombreuses entrées sont toujours présents dans une telle solution. L'ensemble de ces problématiques pousse plutôt à tenter d'élaborer une organisation des données permettant de normaliser l'ensemble des informations obtenues afin d'en simplifier l'accès et l'utilisation.

4.1.2 Schéma décomposé

Une autre forme de stockage disponible se présente sous la forme d'un schéma décomposé dans lequel les informations sont réparties dans différentes tables liées. Cette approche orientée-objet s'adapte bien dans un SGBDR, les relations entre les différents objets pouvant être énoncées formellement. Une telle structure permet de séparer l'ensemble des informations en objets simples, évitant notamment de créer de la redondance et sera par conséquent privilégiée dans le cadre de ce travail.

Une perspective orientée-objet permet en outre de stocker de plus amples informations et de standardiser au maximum les différentes vérifications d'intégrité permettant d'assurer la cohérence des données stockées. Par exemple, les contraintes multiples dans une table unique couvrant les champs station, paramètre, horodatage, voie, direction, vitesse et classe_vehicules pourraient être ramenées simplement aux champs station, paramètre et mesure, les informations de voies et direction étant désormais stockées à l'intérieur même d'un objet station et les informations de classes de vitesses et de véhicules dans un objet paramètre. Les vérifications seraient donc faites sur la relation à un objet et non plus directement sur un ensemble d'informations.

Les relations formelles de dépendance entre les objets permettent aussi de simplifier toutes les modifications, un nombre infini d'objets pouvant être liés à un autre objet spécifique. Une modification des attributs d'une station peut donc être effectuée une seule fois sur un objet station et couvrir l'ensemble des mesures associées par cette relation. Les mesures peuvent aussi aisément être réassignées à une autre station en cas de besoin ou de changement dans l'organisation des données. La relation étant bidirectionnelle, il serait aussi facile de supprimer l'ensemble des mesures liées à une station simplement en supprimant celle-ci.

La structure décomposée, dans une approche de développement itérative, permet plus de versatilité étant donné qu'un bon nombre de modifications peuvent être faites sur des tables complémentaires. Si, par exemple, un nouvel ensemble de données menait à vouloir effectuer des changements dans la structure des objets stations, seulement cette table et ses quelques dizaines ou centaines d'éléments devraient être modifiés plutôt que plusieurs millions d'enregistrements.

Le développement d'un tel schéma demande toutefois une plus grande planification et un travail de design plus poussé, les différents objets devant être répertoriés, en plus de l'ensemble des informations à stocker pour chacun d'entre eux. Les sections suivantes feront état de ce travail de conception et de son implantation dans un ensemble de choix technologiques.

4.2 Les objets répertoriés

Le développement d'un système intégré implique de définir l'ensemble des objets les plus simples afin de décomposer les différents flux de données et d'obtenir des structures communes et de lier ces structures entre elles. Un examen rapide des diverses structures de fichiers recensées au chapitre précédent permet d'identifier 7 objets principaux présentant plusieurs caractéristiques, soit les exploitants de stations, les stations, les paramètres de mesure, les types de mesure, les unités, les mesures et finalement les fichiers d'origine. L'ensemble de ces objets et des différentes propriétés qu'ils peuvent présenter sont détaillés ci-bas.

4.2.1 Exploitant de station

Les données étant fournies par différents organismes, il est intéressant de conserver des traces quant à la provenance des informations et sur les exploitants des stations. Un lien entre chacune des stations et leur propriétaire devrait en outre permettre de les rassembler selon les exploitants lors des phases d'analyse. Ce lien devrait s'appliquer de façon à ce que chaque exploitant soit en mesure de posséder plusieurs stations.

Les autres informations sur cet objet devraient pour l'instant se limiter au nom de l'organisme propriétaire ainsi qu'à des notes supplémentaires, mais pourraient potentiellement

s'enrichir à l'avenir lorsque de nouveaux ensembles de données seront obtenus.

4.2.2 Station

Les stations sont les objets physiques principaux dans le paradigme identifié. Elles représentent les points de collecte de données qui permettent de localiser spatialement un grand nombre de mesures. L'ensemble des stations a comme propriété principale d'appartenir à un exploitant et d'avoir un nombre illimité de mesures qui leur sont attachées, en plus de présenter plusieurs informations qualitatives, telles qu'un nom, un ou plusieurs codes numériques ainsi qu'une position géographique.

Certaines stations, telles que les stations de circulation, présentent des attributs supplémentaires comme un numéro de voie et une direction. Puisque ces informations peuvent être présentes pour seulement un sous-ensemble de données d'une station, il peut être nécessaire d'ajouter un aspect de hiérarchisation en intégrant des objets sous-stations, qui ont comme propriété supplémentaire d'appartenir à une autre station. Dans ces cas particuliers, il s'agit de décomposer une station physique principale en un ensemble de sous-éléments comprenant potentiellement des directions et des voies.

Chaque station peut potentiellement accumuler plusieurs types d'information. Bien que dans les ensembles fournis les stations n'aient en général qu'un seul type de données principal, certaines d'entre elles accumulent tout de même plusieurs types d'informations, notamment sous la forme d'intervalles variables entre les enregistrements ou encore sous la forme de plusieurs sous-types comme les données de qualité de l'air moléculaires et de COV. Par conséquent, les stations ne sont pas associées à des types de données précis et ne devraient pas présenter d'information à ce sujet.

Du point de vue relationnel, les stations, comme mentionné précédemment, appartiennent à des exploitants, et possèdent une quantité indéfinie de mesures.

4.2.3 Type de mesure

Les types de mesures définissent la famille à laquelle appartient une valeur stockée à un point et un moment précis par une station. Dans un système intégré, il est important de pouvoir départager les données selon leur appartenance à une catégorie afin de procéder aux analyses sur un groupe de valeur ayant des propriétés communes. Certaines caractéristiques sont associées au type, notamment l'intervalle de temps sur lequel une mesure a été prise. Ces intervalles étant variables selon les stations, le temps ou encore le paramètre étudié, il est primordial de conserver une telle information, notamment dans le but de procéder à des agrégations et ramener des mesures à intervalles variables à une échelle de résolution

commune afin d'en faire le traitement. Les types de mesure peuvent aussi inclure des sous-types, comme les données de qualité de l'air, qui présentent des mesures pour des paramètres moléculaires ainsi que des COV.

Par conséquent, les types de mesure devraient fournir des informations sur le type et le sous-type de mesure, ainsi que l'intervalle de temps entre les mesures. Les types de mesures sont indépendants et un nombre infini de mesures peuvent dépendre d'un type unique.

4.2.4 Paramètre de mesure

Les paramètres de mesure permettent de connaître la nature des informations accumulées dans la base de données. La décomposition des informations pourrait faire en sorte, par exemple, d'engendrer deux paramètres indépendants pour présenter les caractéristiques du vent, un paramètre étant nécessaire pour présenter la direction du vent, et un autre étant nécessaire pour présenter la vitesse du vent.

Cette structure fondamentale est valable pour l'ensemble des données obtenues, à l'exception des comptages de circulation des stations permanentes qui présentent des niveaux de détails plus importants. Pour ces cas spécifiques, les paramètres doivent aussi inclure la possibilité d'ajouter des informations sur la classe de véhicules ainsi que sur la classe de vitesse des véhicules dans les cas où de telles informations sont fournies. Afin de résoudre ce problème, des champs supplémentaires sont requis exclusivement afin d'inclure ces informations spécifiques aux données de circulation.

Comme dans le cas des types de mesures, les paramètres sont indépendants et peuvent avoir un nombre illimité de mesures qui dépendent de chacun d'entre eux.

4.2.5 Unité

L'unité définit l'échelle de ce qui est mesuré par un enregistrement en particulier. Les mesures peuvent être fournies avec des échelles variables, ce qui implique de séparer les unités des paramètres étudiés.

Les champs devraient inclure des informations permettant d'identifier l'unité elle-même, ainsi qu'une échelle associée dans le but de faire des conversions d'unités lors des analyses si c'est nécessaire. Des exemples d'unités pourraient être $\mu g/m^3$ pour des données moléculaires de qualité de l'air ou plus simplement des véhicules pour les données de comptage de circulation. Les unités sont elles aussi indépendantes, avec un nombre illimité de mesures pouvant dépendre d'une unité unique.

4.2.6 Fichier source

Les fichiers sources ne sont pas des objets nécessaires à l'interprétation des données fournies, mais sont essentiels pour mener à bien les efforts d'assainissement de la base de données. La conservation des informations quant à la provenance de chacune des mesures devrait permettre d'éventuellement réparer des erreurs associées à des fichiers spécifiques si de tels cas devaient se présenter. La présence d'informations liées au fichier devrait aussi permettre de gérer des conflits potentiels avec de nouveaux ensembles de fichiers qui pourraient présenter des chevauchements de périodes d'analyse et ainsi comprendre des données déjà présentes dans la base de données.

Chaque instance d'un fichier source devrait conserver des renseignements quant au nom de fichier, le moment de l'insertion dans la base de données, l'utilisateur ayant procédé à l'insertion, ainsi que le nombre de mesures associées. Les relations des fichiers sources se limitent à l'association de chaque mesure à un fichier d'origine spécifique.

4.2.7 Mesure

Les mesures sont le point central de ce qui deviendra le schéma de données. Tous les objets énoncés plus tôt, à l'exception des exploitants de stations, sont des dépendances directes des mesures, qui ne peuvent exister sans les informations incluses dans ces objets.

Les objets mesure entreposent cependant l'information la plus névralgique, soit la valeur de la mesure effectuée, ainsi que le moment auquel celle-ci a été prise. Dans certains cas, des informations quant à la qualité de la mesure pourraient être fournies, et à cet effet, un espace devrait être réservé pour chaque mesure pour pourvoir à cette éventualité.

4.3 Élaboration d'un schéma de données

La définition d'un schéma de données consiste à reprendre l'ensemble des objets répertoriés et leurs caractéristiques et à définir des champs permettant de faire le stockage de chacun d'entre eux. Cette structure doit aussi s'accompagner des mesures mises en place afin d'assurer la validité des données qui y sont stockées sous la forme d'un ensemble de contraintes, ainsi qu'un certain nombre de normes permettant d'organiser efficacement les données et d'assurer la validité des contraintes définies.

4.3.1 Le schéma

Suite à l'identification des objets, il est désormais possible de préciser le contenu de chacun de ceux-ci et d'expliciter les relations entre chacun. La Figure 4.1 offre une représentation

graphique de chacun des objets [1], des champs qu'ils devraient contenir afin d'entreposer l'ensemble des informations liées aux objets, ainsi que de la nature de chacun de ces champs.

Les Tableaux 4.4, 4.5, 4.6, 4.7, 4.8, 4.9 et 4.10 font quant à eux la description des objets, des champs et des informations qui devraient y être entreposées.

Tableau 4.4: Structure de l'objet source

Champ	Type de données	Information entreposée
id	Entier auto-incrémenté	Numéro d'identifiant unique
nom	Chaîne de caractère	Le nom de l'exploitant de la station
note	Texte	Informations supplémentaires sur l'exploitant

1. Dans un souci de simplification, les fichiers source sont rassemblés dans l'objet lot, alors que les exploitants de stations sont rassemblés dans l'objet source.

Tableau 4.5: Structure de l'objet station

Champ	Type de données	Information entreposée
id	Entier auto-incrémenté	Numéro d'identifiant unique
code	Chaîne de caractères	Le code numérique principal de la station
nom	Chaîne de caractères	Le nom de la station
nom_court	Chaîne de caractères	Un nom plus approprié pour la présentation des résultats
source_id	Entier	Référence à un identifiant unique de l'objet source
alias	Chaîne de caractères	Pseudonymes de la station
position	Référence géographique	Une référence géographique de la position de la station
latitude	Nombre à virgule	La latitude de la station
longitude	Nombre à virgule	La longitude de la station
direction	Chaîne de caractères	La direction dans laquelle le mesure est effectuée (circulation seulement) [2]
voie	Chaîne de caractères	La voie sur laquelle la mesure est effectuée (circulation seulement) [3]
parent_id	Entier	Référence à l'identifiant unique d'une autre station (dans les cas ou une station est une sous-station)

Tableau 4.6: Structure de l'objet type

Champ	Type de données	Information entreposée
id	Entier auto-incrémenté	Numéro d'identifiant unique
nom	Chaîne de caractères	Le nom du type de données
sous_type	Chaîne de caractères	Le sous-type (si nécessaire)
intervalle	Entier	Le nombre de secondes entre les mesures

2. Dans le cas des carrefours, la direction est utilisée pour indiquer la destination des véhicules.
3. Dans le cas des carrefours, la voie est utilisée pour indiquer l'origine des véhicules.

Tableau 4.7: Structure de l'objet paramètre

Champ	Type de données	Information entreposée
id	Entier auto-incrémenté	Numéro d'identifiant unique
nom	Chaîne de caractères	Le nom du paramètre [4]
nom_html	Chaîne de caractères	Le nom du paramètre pour l'impression à l'écran en format html
valeur_min	Nombre à virgule	La valeur minimale de la classe [5]
valeur_max	Nombre à virgule	La valeur maximale de la classe

Tableau 4.8: Structure de l'objet unite

Champ	Type de données	Information entreposée
id	Entier auto-incrémenté	Numéro d'identifiant unique
nom	Chaîne de caractères	L'unité utilisée
nom_html	Chaîne de caractères	Le nom de l'unité pour l'impression à l'écran en format html
multiple	Nombre à virgule	Une valeur pouvant servir à faire la conversion d'une unité vers une autre

Tableau 4.9: Structure de l'objet lot

Champ	Type de données	Information entreposée
id	Entier auto-incrémenté	Numéro d'identifiant unique
usager	Chaîne de caractères	Le nom de l'usager ayant procédé à l'insertion du lot
horodatage	Horodatage	La date et l'heure de l'insertion du lot
fichiers_source	Chaîne de caractères	Le(s) nom(s) du/des fichier(s) présent(s) dans le lot
nombre_entrees	Entier	Le nombre de données associées à ce lot

4. La classe de véhicules pour les données de circulation détaillées.

5. Au moment d'écrire ces lignes, les champs valeur_min et valeur_max ne sont utilisés que pour les classes de vitesse des données de circulation.

Tableau 4.10: Structure de l'objet mesure

Champ	Type de données	Information entreposée
id	Entier auto-incrémenté	Numéro d'identifiant unique
station_id	Entier	Référence à un identifiant unique de l'objet station
parametre_id	Entier	Référence à un identifiant unique de l'objet parametre
horodatage	Horodatage	La date et l'heure de la prise de mesure (fin de la période)
valeur	Nombre à virgule	La valeur de la mesure
qualite	Chaîne de caractères	La qualité de la mesure enregistrée
type_id	Entier	Référence à un identifiant unique de l'objet type
unite_id	Entier	Référence à un identifiant unique de l'objet unite
lot_id	Entier	Référence à un identifiant unique de l'objet lot

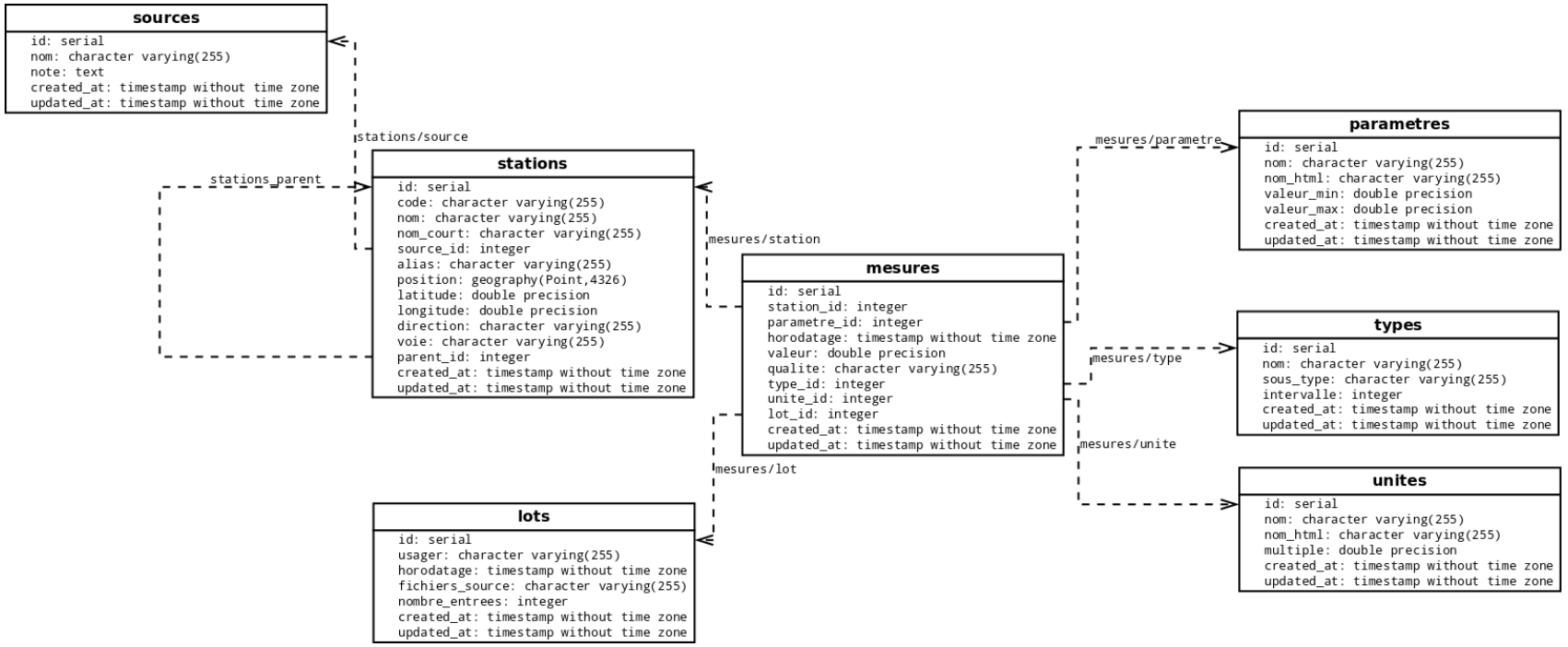


Figure 4.1: Structure complète du schéma défini

4.3.2 Contraintes d'insertion

Bien que le schéma lui-même soit défini, et qu'ainsi l'ensemble des informations à y insérer ait un point de stockage précis, il importe d'instaurer des règles de contrôle permettant de garantir la cohérence des entrées qui y seront faites. Cet ensemble de contraintes, sous la forme de clés primaires, devrait permettre d'empêcher notamment l'insertion de tous les doublons potentiels identifiés au Chapitre 3. Chacun des objets énoncés précédemment doit faire l'objet de vérifications spécifiques afin d'assurer l'unicité des informations stockées dans la base de données.

Certains des objets présentent des caractéristiques d'unicité simple selon un seul champ, soit des clés primaires simples. Les sources devraient faire l'objet d'un contrôle afin d'éviter qu'un même exploitant de station présente plus d'une seule entrée. Les lots devraient quant à eux être uniques selon leur fichier d'origine, offrant ainsi une première protection contre l'entrée de valeurs doubles.

Pour les autres cas, l'unicité doit être établie sur plusieurs champs, soit des clés primaires composées. Les paramètres doivent en effet ne présenter qu'une seule combinaison de nom, valeur_min et valeur_max. Les différentes classes de véhicules des données de circulation pourront alors présenter plusieurs entrées, avec des classes de vitesses différentes, alors que les autres ensembles resteront uniques pour chaque valeur de nom, ceux-ci ne présentant pas d'informations où le paramètre présente un intervalle borné par des bornes inférieure et supérieure.

Les types devraient quant à eux être contraints selon leurs champs nom, sous_type et intervalle. Ainsi, les différents types pourront être associés à plus d'un intervalle de temps, permettant ainsi aux mesures d'être associées à leur type approprié. Les unités doivent de leur côté être vérifiées de telle sorte que chaque nom soit unique.

Les stations doivent avoir leur unicité vérifiée en fonction de leur exploitant, d'une direction, d'une voie ainsi que d'un parent. Une telle vérification permettra d'éviter de créer plus d'une sous-station présentant les informations pour une même voie ou direction.

Le cas des mesures est probablement le plus complexe à gérer, étant donnée sa position de dépendance envers plusieurs autres tables. Les mesures doivent en effet être uniques selon leur station d'origine, leur paramètre ainsi que selon leur horodatage. Ainsi, une seule mesure pourra exister pour un site spécifique et pour un paramètre précis, à une heure donnée. Cette façon de procéder est en accord avec la façon de faire définie préalablement où les données de circulation sont associées à des sous-stations selon la voie et la direction. Il pourrait sembler intéressant d'ajouter le type à cette validation, mais ceci présente le défaut de permettre d'insérer plusieurs mesures à des échelles différentes à la même heure, une superposition qui pourrait entraîner des inexactitudes si les entrées ne sont pas vérifiées individuellement lors

de l'analyse. Pour cette raison, le choix est fait de ne pas permettre plus d'un intervalle à une heure précise, et de favoriser les données les plus désagrégées disponibles. L'obtention des données sur des intervalles plus grands pourra se faire simplement en faisant des agrégations. Finalement, des vérifications quant aux unités ne se révèlent pas non plus pertinentes, étant donné que de simples conversions devraient permettre d'obtenir l'information selon l'unité de mesure désirée.

4.3.3 Normalisation des données

Les ensembles de données fournis possèdent des normes de présentation de l'information très variables en fonction des types et des différents formats de données. Afin de procéder à la mise en commun, il est nécessaire d'établir certaines conventions afin de simplifier le traitement ultérieur des données. L'établissement de normes strictes est aussi névralgique dans le but que l'ensemble des contraintes définies puisse être appliqué correctement.

Horodatages

L'information faisant le plus l'objet de variations entre les ensembles de données fournies est la codification du moment de la prise de mesure et l'intervalle sur lequel la mesure s'applique. Certains des ensembles de données, notamment pour la qualité de l'air et la météo, utilisent une codification de fin d'intervalle de mesure, alors que les données de circulation utilisent le début de l'intervalle de mesure pour référencer l'élément mesuré. De telles variations ne peuvent pas coexister dans un système intégré, et il importe donc d'établir une norme fédératrice pour l'ensemble de toutes les mesures.

Pour des raisons principalement de logique, notamment parce qu'une mesure horaire ne peut pas être obtenue avant la fin la période de mesure, la fin de l'intervalle de temps sur lequel la mesure s'applique est choisie comme norme. Toutefois, pour des raisons pratiques, notamment afin que les données soient associées à la bonne journée dans le cas des données sur le dernier intervalle de la journée, le temps de fin de période est ramené d'une seconde. Par conséquent, une mesure couvrant l'intervalle de 15:00 à 16:00 sera codifiée dans la base de données comme 15:59:59. Cette codification permet d'accéder à l'ensemble des mesures sur une heure en faisant une requête ouverte dans la base de données sur toutes les valeurs commençant par "15:", en plus d'associer toutes les valeurs aux jours appropriés.

Un autre facteur à prendre en considération dans la gestion du temps est celui des fuseaux horaires. En effet, puisque certains ensembles de données (Qualité de l'air, Météo) ne tiennent pas compte des changements d'heures aux mois de mars et novembre de chaque année, alors que les comptages de véhicules utilisent l'heure civile, il est essentiel de normaliser selon

un fuseau horaire commun. Étant donné que la plupart des analyses doivent être faites à l'heure civile, le choix d'utiliser à la fois l'heure normale de l'Est et l'heure avancée de l'Est en fonction de la date se révèle plus simple pour les phases d'analyse. Par conséquent, toutes les lectures de qualité de l'air et de météo faites entre le second dimanche de mars à 2:00 et le premier dimanche de novembre à 2:00 devront voir leurs horodatages être additionnés d'une heure afin de respecter cette heure civile.

Finalement, un dernier objet de normalisation pour les horodatages est la présence de données manquantes dans les ensembles de données. Comme stipulé à la Section 3.3, certaines données ne retiennent pas la notion d'une entrée invalide et par conséquent un bon nombre d'heures sont manquantes, principalement dans le cas des données de la qualité de l'air. Dans ce cas précis, les données manquantes se révèlent particulièrement importantes compte tenu du besoin de calculer des moyennes mobiles sur un certain nombre d'heures sur lequel un seuil de qualité de l'air s'applique avec une proportion minimale de mesures. Encore ici, pour simplifier les calculs dans les phases d'analyses, le choix est fait de procéder à l'insertion de valeurs nulles pour les cas où des données sont manquantes.

Hiérarchisation des stations

Les Sections 4.2 et 4.3.1 font état du besoin de hiérarchiser les stations en raison de la complexité de celles-ci dans le cas des données de circulation. Cette séparation des mesures en plusieurs sous-stations et le besoin lors des étapes d'analyses de faire la somme de celles-ci sur différents niveaux de résolution, par exemple pour une seule direction ou pour les deux directions à un site de comptage donné, demande l'imposition d'une norme stricte afin que des processus automatisés puissent présenter l'information à ces échelles. La norme définie devrait aussi permettre d'inclure des variations dans la géométrie des axes routiers mesurés, comme l'ajout de voies ou la transformation d'une artère à sens unique en route bidirectionnelle, sans devoir effectuer de modifications dans la méthode de calcul.

Pour toutes ces raisons, une norme permettant trois niveaux hiérarchiques de station est choisie. Par conséquent, une station principale peut être décomposée en une ou plusieurs sous-stations représentant des directions, alors que ces sous-stations peuvent elles aussi avoir des sous-stations représentant des voies physiques leur étant rattachées. Une représentation graphique illustrant cette hiérarchisation dans le cas des données détaillées sur les ponts est présentée à la Figure 4.2. La station principale (orange) possède deux sous-stations en direction nord et sud (bleu), qui elles-même possèdent 6 sous-stations (vert) faisant le lien avec les mesures. Dans le cas d'agrégations, ou si uniquement des mesures agrégées à un niveau supérieur devaient être fournies dans un envoi subséquent, celles-ci pourraient être associées indépendamment aux sous-stations appropriées. Il va sans dire qu'il est toutefois nécessaire

d'effectuer des validations afin de ne pas procéder à l'insertion de données à des niveaux d'agrégation différents pour les mêmes périodes, puisqu'une telle superposition entraînerait des valeurs supérieures aux mesures réelles.

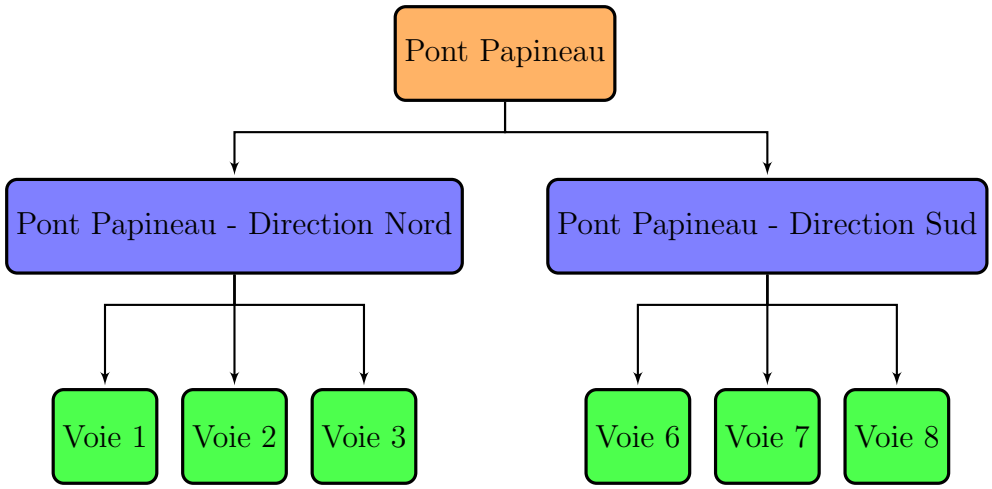


Figure 4.2: Hiérarchie des stations et sous-stations du pont Papineau

La hiérarchisation s'effectue de façon quelque peu semblable pour un site où la géométrie actuelle ne permet qu'une seule direction, comme dans le cas d'un comptage sur une seule direction sur une autoroute. Un schéma illustrant la structure se présente à la Figure 4.3. Dans ce cas particulier, un seul niveau existe, soit une station parent (bleu). La structure permet toutefois d'ajouter de nouveaux comptages pour chacune des voies à l'avenir si jamais celles-ci existent et devenaient disponibles (vert pâle). Par ailleurs, même si dans ce cas particulier il est improbable que la géométrie change de façon aussi radicale, la hiérarchie permet de créer une station principale (orange) dont pourrait dépendre par la suite un autre ensemble de sous-stations (sous-arbre de droite). Même si de telles variations sont improbables et limitées à quelques cas potentiels, il est important de définir une norme permettant de réorganiser aisément les relations entre les stations et ainsi permettre l'extensibilité du système. Finalement, afin d'assurer la compatibilité entre les sites de comptages présentant des arbres complets avec ceux présentant des arbres partiels, les stations au plus haut des arbres se verront attribuer une référence à elles-mêmes dans le champ "parent_id", permettant ainsi de toujours pouvoir faire des regroupements selon ce champ.

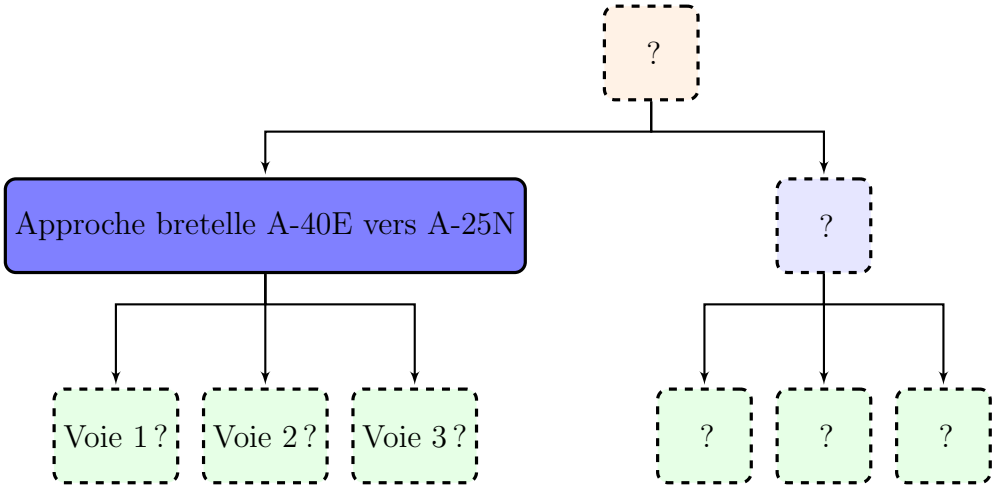


Figure 4.3: Hiérarchie des stations et des sous-stations pour l'approche de la bretelle de l'A-40E vers l'A-25N

Localisation géographique

L'utilisation de différents systèmes de coordonnées peut rendre la tâche d'identifier clairement le positionnement d'une station complexe, notamment dans le cas de systèmes projetés. Afin d'éviter des risques potentiels, la façon de faire la plus simple est de localiser l'ensemble des stations dans un système utilisant les valeurs de latitude et de longitude. Le standard le plus commun pour de telles valeurs, soit le système géodésique WGS84, notamment utilisé pour la codification des données GPS, a été choisi pour codifier l'ensemble des coordonnées géographiques à intégrer au système d'information. La plupart des logiciels modernes permettant d'arrimer des données géographiques à des SGBDR peuvent gérer directement de telles localisations et appliquer un bon nombre de fonctions directement sur celles-ci et ainsi limiter les risques d'erreurs de saisie par l'utilisateur (de tels outils seront abordés à la Section 4.4).

4.4 Technologies utilisées et implantation

Les différentes sections précédentes font état des étapes de conception définissant l'architecture de l'entrepôt de données qui sera utilisé. Il importe désormais de sélectionner des technologies permettant d'implanter le schéma ainsi que les ensembles de contraintes définis. Afin d'y parvenir, il est pertinent de formellement énoncer les différents besoins liés au projet, avant d'étudier un ensemble de technologies pouvant y répondre. Une fois ces besoins énoncés et des technologies pouvant les combler choisies, l'implantation du schéma à l'intérieur de ces dernières sera décrite.

4.4.1 Besoins et choix technologiques

Afin de réaliser le montage d'un système d'information qui pourra soutenir efficacement l'analyse, il importe de cerner un ensemble de besoins afin que celui-ci puisse répondre efficacement aux attentes. Des technologies permettant de combler ces attentes doivent par la suite être choisies.

Besoins

Les travaux réalisés jusqu'à maintenant permettent d'énumérer un ensemble de besoins primaires à la réussite du projet.

Premièrement, la taille des ensembles de données décrits précédemment au Chapitre 3 fait en sorte que les solutions logicielles devraient avoir la capacité d'emmagasiner des quantités importantes d'informations sans voir leurs performances diminuer de façon considérable. De plus, le système de gestion de données devrait pouvoir permettre à plusieurs utilisateurs de faire un usage constant des données entreposées. Les aspects relationnels entre les objets mentionnés précédemment, ainsi que ces deux besoins spécifiques, font porter le choix vers un SGBDR utilisant une relation serveur-client, ce qui devrait permettre un entreposage centralisé et un accès distant pour tous les travaux subséquents à ceux énoncés dans le cadre de ce mémoire. Finalement, le SGBDR devrait idéalement pouvoir gérer des objets géographiques, permettant de réaliser différentes analyses spatiales.

Les disparités recensées entre les différents types et ensembles de données ainsi que l'obligation de faire des traitements sur presque tous les enregistrements à insérer entraînent le besoin de faire l'utilisation d'un langage de programmation de haut niveau permettant de faire les manipulations nécessaires. Afin de conserver l'approche orientée-objet déjà développée dans la création du schéma, ce langage devrait idéalement appartenir à ce paradigme de programmation, en étant le plus possible modulaire, permettant ainsi une réalisation du projet dans un processus itératif selon la méthode de développement Agile. Évidemment, la capacité du langage de programmation à communiquer avec le SGBDR choisi est un aspect fondamental qui devra être pris en compte lors du choix de celui-ci. La possibilité pour le langage de programmation retenu d'intégrer des greffons et bibliothèques externes devrait aussi être retenue comme un critère de sélection dans le but d'assurer la viabilité à long terme du système d'information. Finalement, le langage choisi devrait offrir des options dans le but de produire des résultats de façon graphique, et idéalement de façon interactive et à distance.

Choix technologiques

Les SGBDR pouvant répondre aux besoins sont nombreux, les plus connus dans le monde du logiciel libre [6] étant PostgreSQL, MySQL et SQLite. Bien qu'il réponde à la plupart des critères, notamment en incluant des capacités géographiques, SQLite doit être rejeté étant donné qu'il n'inclut pas de serveur permettant un accès à plusieurs usagers, l'ensemble d'une base de données étant conservée dans un fichier unique. PostgreSQL et MySQL jouissent tous deux de bonnes réputations quant à leur capacité à gérer de très grandes bases de données. Les deux SGBDR possèdent aussi des fonctions internes afin de générer des vues et des vues matérialisées [7] nécessaires compte tenu des besoins de présenter les données à différents niveaux de résolution. L'obtention de références sérieuses permettant de les départager et de décrire un des deux logiciels comme étant plus apte que l'autre à gérer les grands ensembles de données est toutefois difficile, et hors du cadre de ce travail. PostgreSQL présente toutefois un avantage marqué grâce aux extensions de PostGIS, qui incluent un grand nombre de fonctions géographiques. Pour cette raison, le choix a été fait de baser le système d'information sur le SGBDR PostgreSQL et ses extensions PostGIS.

Les différentes alternatives répondant aux besoins mentionnés précédemment en ce qui a trait au langage de programmation abondent, ce qui fait en sorte que le processus peut se révéler fastidieux. Afin de le simplifier, les recherches pour le langage de programmation se porteront uniquement sur des langages de haut niveau interprétés. Ceux-ci ont comme avantage de pouvoir être utilisés sur un ensemble de plates-formes différentes, en plus de faciliter le développement en raison de leur nature dynamique. Ces avantages se font au prix d'une perte de performance des différents scripts développés. Ce coût en performance peut toutefois être considéré comme acceptable, les processus les plus lourds en terme de ressources matérielles étant liés à l'insertion des données dans le SGBDR. L'exécution de tels processus devant uniquement être complétée lorsque de nouveaux ensembles de données sont rendus disponibles, les pertes en termes de temps de calcul devraient s'avérer très limitées.

Les langages interprétés orientés-objet les plus connus sont PHP, Python, Ruby ou encore Javascript. PHP et Javascript sont principalement utilisés dans le développement de site et d'interface web, alors que Python et Ruby sont multi-usages et sont la fondation de nombreux logiciels. La versatilité de Python et de Ruby les rend tous deux plus intéressants, ceux-ci pouvant être exécutés directement sans devoir passer par un serveur ou un navigateur web. En outre, ils possèdent tous deux des écosystèmes actifs, Python étant utilisé dans un plus grand nombre de projets, et peuvent se connecter sans problèmes à un serveur PostgreSQL.

6. Les logiciels libres ont été préférés dans le cadre de ce travail en raison de leur capacité à répondre aux besoins et des coûts associés aux systèmes propriétaires équivalents.

7. Une vue matérialisée permet de conserver physiquement les informations contenues dans une vue afin de permettre un accès rapide à celles-ci.

Afin de compléter cette connexion au SGBDR, les deux langages possèdent des “frameworks” offrant une couche d’abstraction de base de données et un Mapping objet-relationnel (MOR) permettant de traiter chaque entrée de base de données comme un objet. Les deux “frameworks” les plus évolués sont respectivement Django pour Python et Ruby on Rails pour Ruby [8]. Finalement, ils possèdent tous deux les greffons nécessaires afin d’utiliser les capacités géographiques des extensions PostGIS.

Les deux solutions se révèlent donc plutôt similaires et aptes à répondre aux besoins du système d’information. Ruby et son framework Ruby on Rails ont été choisis, les capacités du système à gérer les fuseaux horaires étant mieux intégrés, un besoin important compte tenu de la nécessaire gestion des changements de l’heure normale à l’heure avancée pour certains ensembles de données.

Rails permet de générer aisément des modifications à la structure de la base de données grâce à l’utilisation de migrations, soit un ensemble de fichiers cumulatifs qui contrôlent le schéma des différentes tables. Ces fichiers ne sont exécutés qu’à une seule reprise et en séquence, et contiennent un ensemble d’instructions à exécuter sur le schéma. Alors que les migrations contrôlent la structure de la base de données, Rails intègre un autre concept, soit les modèles, qui permettent de définir différents types d’objets à l’intérieur d’un projet. Un modèle doit donc être associé à chaque table de la base de données afin de générer des objets pour chacune d’entre elles. L’utilisation de modèles permet aussi un accès direct aux vues ou vues matérialisées, ce qui fait en sorte que les données peuvent être accédées directement sans égards à leur provenance ou aux manipulations internes requises afin de les obtenir.

Ruby et Rails peuvent aussi faire l’usage de greffons, appelés “gem”, qui viennent étendre les capacités du langage et fournir des méthodes qui seront utiles dans le cadre de ce projet. Finalement, Rails étant principalement utilisé dans le développement de sites web dynamiques, la génération d’outils visuels interactifs se fera sans problème.

4.4.2 Implantation du système d’information

Le choix de baser le système sur Rails et PostgreSQL étant fait, il est possible de procéder à l’implantation des structures définies afin de stocker les données. Cette implantation du schéma et des contraintes à l’intérieur d’un projet Rails se fait aisément, sur la base d’un ensemble de fichiers de migrations et de modèles.

8. L’appellation Rails, plus simple, sera utilisée dans le reste de cet ouvrage.

Migrations

Les fichiers de migrations, tel qu'indiqué précédemment, présentent des indications quant à la structure de la base de données. Puisqu'ils sont exécutés en séquence, il est possible d'y ajouter ou de modifier les champs de façon graduelle si le schéma présenté à la Section 4.3 doit être modifié. La nature séquentielle du processus de création de la structure fait en sorte que les dépendances d'une table doivent être créées préalablement à celle-ci. Ces relations peuvent toutefois être modifiées par une nouvelle modification si un changement dans la structure du schéma le requiert. Par conséquent, la table sources doit être créée avant la table stations, et les tables stations, types, paramètres, unités et lots doivent être créées avant la table mesures. Dans le but de présenter simplement la création des tables de la base de données, l'ensemble des migrations a été réduit à un seul fichier par objet identifié, pour un total de 7 migrations.

La structure des fichiers se révèle assez simple, la première ligne indiquant que le fichier s'ajoute aux migrations sous la forme suivante "class CreateNomDeLaClasse < ActiveRecord::Migration". Par la suite, les changements à faire à la structure de la base de données sont définis dans une méthode nommé "change" et se résument, dans les cas présentés ici, à la création de nouvelles tables. Le reste du fichier fait état des champs à créer, en y exposant notamment leur type, leur nom ainsi que des informations complémentaires.

Les types disponibles comportent notamment les "string", "text", "float", "integer", "timestamp" et "references", créant respectivement des champs de chaînes de caractères, de texte, de nombres à virgule, de nombres entiers, d'horodatage ou de nombre entier faisant référence à d'autres champs. L'ensemble des types disponibles peut être consulté dans la documentation de Rails. Il est à noter que des gems peuvent venir enrichir les types disponibles par défaut, ce qui sera le cas ici notamment lors de la création de champs comportant des objets géographiques PostGIS.

Les fichiers de migrations comportent aussi d'autres informations, notamment en ce qui a trait à la création d'index. Ceux-ci permettent de référencer certaines colonnes afin d'accélérer les requêtes sur celles-ci. Ces index peuvent aussi servir à établir les contraintes d'unicité au niveau de la base de données, une façon de procéder qui est plus efficace que les vérifications d'unicité interne de Rails (voir à ce sujet la section suivante).

Finalement, l'utilisation du gem "foreigner" permet de définir les relations entre les tables directement dans la base de données sous la forme de clés étrangères, alors que les relations entre les objets sont normalement définies dans les modèles dans l'environnement Rails. Ces relations ne sont normalement pas référencées directement dans la base de données, Rails ayant la possibilité de se connecter à certains SGBDR ne possédant pas les capacités nécessaires. Cette définition de clés étrangères s'avère toutefois importante dans le cadre de ce projet, étant donné que les données devront parfois être accédées à l'aide de logiciels externes

comme un logiciel de Système d'Information Géographique (SIG), et que ceux-ci n'ont pas la capacité de reconnaître les relations établies dans les modèles de rails.

Le Code 4.1 présente la structure du fichier migration pour les exploitants de stations (sources). La première ligne vient définir l'espace dans lequel la migration s'effectue, alors que la ligne 2 spécifie la méthode à utiliser, qui sera toujours "change" dans le cas des migrations créant une table. Les deux champs définis préalablement, soit "nom" et "note", sont référencés aux lignes 4 et 5, comme étant des champs de la table "sources" tel que spécifié à la ligne 3. Rails ajoute automatiquement à toutes les tables les champs id d'identifiant unique, ainsi que des champs horodatages permettant de conserver des renseignements sur le moment de la création de l'entrée et sur le moment de la dernière mise à jour de l'entrée, qui sont spécifiés à la ligne 7 avec le type de données "timestamps".

Code 4.1: Migration créant la table sources

```
1 class CreateSources < ActiveRecord::Migration
2   def change
3     create_table :sources do |t|
4       t.string :nom, :null => false
5       t.text :note
6
7       t.timestamps
8     end
9   end
10 end
```

La table stations présente une plus grande quantité de champs, et inclut deux éléments relationnels, ce qui fait en sorte que la migration la créant est plus longue. Les lignes 4 à 14 du Code 4.2 présentent l'ensemble des champs ayant été décrétés plus tôt comme faisant partie de l'objet station. La ligne 4 stipule qu'un code d'identification doit obligatoirement être fourni pour le champ, forçant ainsi le stockage d'au moins un nom pour chaque objet. Les extensions PostGIS permettent d'ajouter des objets géographiques, ce qui est fait à la ligne 9 à l'aide d'un champ de type point dans le cas des stations.

La ligne 18 établit les index pertinents pour la table stations. La taille anticipée de la table stations ne justifie toutefois pas d'ajouter de tels index sur un grand nombre de champs, et seulement la référence à une source est indexée.

Finalement, les ligne 19 et 20 prennent en charge la création de relations directes entre les tables sous la forme de clés étrangères. Dans le cas de stations, ces clés sont créées avec les sources, ainsi qu'avec la table stations elle-même entre le champ "parent_id" et "id", qui

établissent les relations hiérarchiques entre les stations et leurs sous-stations.

Code 4.2: Migration créant la table stations

```
1 class CreateStations < ActiveRecord::Migration
2   def change
3     create_table :stations do |t|
4       t.string :code, :null => false
5       t.string :nom, :null => false
6       t.string :nom_court
7       t.references :source, :null => false
8       t.string :alias
9       t.point :position, :geographic => true
10      t.float :latitude
11      t.float :longitude
12      t.string :direction
13      t.string :voie
14      t.references :parent
15
16      t.timestamps
17    end
18    add_index :stations, :source_id
19    add_foreign_key(:stations, :sources)
20    add_foreign_key(:stations, :stations, :column => 'parent_id')
21  end
22 end
```

La table types est beaucoup plus simple, ne comprenant que trois champs et l'utilisation d'un index n'y est pas nécessaire. Le Code 4.3 présente bien les trois champs "nom", "sous_type" et "intervalle", avec pour seule contrainte que toutes les entrées doivent posséder au minimum un nom. Cette situation s'explique par le fait que toutes les mesures ne possèdent pas des sous-types et par le fait que certaines données n'incluant pas la notion d'intervalles pourraient éventuellement être ajoutées.

Code 4.3: Migration créant la table types

```
1 class CreateTypes < ActiveRecord::Migration
2   def change
3     create_table :types do |t|
4       t.string :nom, :null => false
5       t.string :sous_type
6       t.integer :intervalle
7
8       t.timestamps
9     end
10  end
11 end
```

Comme pour les types, la définition de la table paramètres se révèle assez simple, avec pour seule contrainte l'insertion d'un nom non nul. Le Code 4.4 présente les commandes menant à la création de la table. Tel que stipulé précédemment, les champs "valeur_min" et "valeur_max" sont utilisés respectivement pour les bornes inférieures et supérieures des classes de vitesse dans le cas des données de circulation uniquement et par conséquent aucune contrainte n'y est appliquée.

Code 4.4: Migration créant la table paramètres

```
1 class CreateParametres < ActiveRecord::Migration
2   def change
3     create_table :parametres do |t|
4       t.string :nom, :null => false
5       t.string :nom_html
6       t.float :valeur_min
7       t.float :valeur_max
8
9       t.timestamps
10    end
11  end
12 end
```

Les unités représentent un autre cas simple ou seul un nom est requis pour la création d'une entrée. Le Code 4.5 présente l'ensemble des instructions fournies à Rails pour la création de la table.

Code 4.5: Migration créant la table unités

```
class CreateUnites < ActiveRecord::Migration
  def change
    create_table :unites do |t|
      t.string :nom, :null => false
      t.string :nom_html
      t.float :multiple

      t.timestamps
    end
  end
end
```

La table lots présente un plus grand nombre de contraintes que les tables précédentes. En effet, afin de bien entreposer l'historique de l'insertion des données dans le système d'information, un nom d'utilisateur, la date et l'heure de l'insertion ainsi que les fichiers de provenance doivent être fournis. Le nombre d'entrées insérées ne pouvant être ajouté qu'après l'insertion des données, l'obligation de fournir une information empêcherait de créer des lots avant l'exécution d'un processus d'insertion, ce qui poserait problème si jamais l'opération devait être interrompue, puisqu'il serait alors impossible de supprimer les mesures déjà insérées. Par conséquent, aucune contrainte n'est appliquée sur le champ.

Code 4.6: Migration créant la table lots

```
class CreateLots < ActiveRecord::Migration
  def change
    create_table :lots do |t|
      t.string :usager, :null => false
      t.timestamp :horodatage, :null => false
      t.string :fichiers_source, :null => false
      t.integer :nombre_entrees

      t.timestamps
    end
  end
end
```

La table mesures présente le plus de contraintes, qui sont présentées au Code 4.7. Tout d'abord, il est essentiel de fournir un ensemble de références s'appliquant à la valeur puisque

celle-ci ne présente aucune information qualificative si les liens avec les autres tables ne sont pas forcés. De plus, la présence d'un horodatage est aussi essentielle puisque les mesures doivent être définies à des moments précis dans le temps.

Par ailleurs, puisque la table mesures devrait à terme être le point central de l'ensemble de toutes les données et ainsi contenir plusieurs millions d'entrées, il est essentiel de s'attarder avec attention à la création d'index permettant des accès à ces données à des vitesses acceptables. L'indexation des champs conservant les informations quant aux stations, paramètres, types, unités, lots et horodatage devraient couvrir les principales variables selon lesquelles les requêtes sur les mesures devraient être effectuées.

La ligne 19 fait état d'une contrainte d'unicité mise en place de façon interne à la base de données et stipule qu'une mesure devrait être unique selon la station d'origine, un paramètre ainsi qu'une heure de mesure. La plupart des autres tables comportent aussi des contraintes à l'insertion implantées dans les modèles afin d'assurer la cohérence de la base de données, mais ces contraintes sont en général simples et ne s'appliquent qu'à un nombre d'entrées limité et peuvent être définies et modifiées plus facilement dans les modèles. Or, les contraintes implantées dans les modèles sont beaucoup plus coûteuses en temps de calculs que les vérifications internes de PostgreSQL, en particulier pour les grands ensembles de données, ce qui explique le choix de procéder à l'implantation d'une contrainte directement dans le SGBDR.

Finalement, la table mesures étant le point central faisant le lien entre toutes les autres, il est nécessaire d'y établir, comme pour la relation entre les stations et les sources, des clés étrangères permettant d'officialiser au niveau de la base de données les relations avec les autres objets définis.

Code 4.7: Migration créant la table mesures

```
1 class CreateMesures < ActiveRecord::Migration
2   def change
3     create_table :mesures do |t|
4       t.references :station, :null => false
5       t.references :parametre, :null => false
6       t.timestamp :horodatage, :null => false
7       t.float :valeur
8       t.string :qualite
9       t.references :type, :null => false
10      t.references :unite, :null => false
11      t.references :lot, :null => false
12
13      t.timestamps
14    end
15    add_index :mesures, :station_id
16    add_index :mesures, :parametre_id
17    add_index :mesures, :type_id
18    add_index :mesures, :unite_id
19    add_index :mesures, ["station_id", "parametre_id", "horodatage"], :unique =>
          true
20    add_index :mesures, :lot_id
21    add_index :mesures, :horodatage
22    add_foreign_key(:mesures, :stations)
23    add_foreign_key(:mesures, :parametres)
24    add_foreign_key(:mesures, :unites)
25    add_foreign_key(:mesures, :types)
26    add_foreign_key(:mesures, :lots)
27  end
28 end
```

Modèles

Les modèles contiennent les informations relatives aux objets et font le lien avec la structure de la base de données afin que les entrées des tables puissent être utilisées directement comme des objets dans le cadre d'un projet Rails. Chaque objet a donc un fichier de modèle associé comprenant les diverses définitions, ainsi qu'un ensemble de méthodes associées, pour un total de 7 fichiers dans le cas des objets définis précédemment.

La structure générale d'un fichier de modèle comprend une première ligne identifiant où l'objet se place dans la hiérarchie des objets internes de Ruby et de Rails. Dans le cas des

objets liés à une table de base de données, une telle définition se fait dans la première ligne de chaque fichier. Il est par la suite nécessaire de définir un certain nombre d'attributs qui seront publiquement accessibles, ceux-ci représentant des champs dans les différentes tables définies.

Les modèles de Rails se chargent aussi d'établir les relations entre les différents objets pour un usage interne. Afin de définir les relations, un ensemble de mots-clés peuvent être utilisés. Si un objet possède plusieurs instances d'un autre objet, la commande "has_many" permettra d'établir la relation, ce qui créera une méthode permettant d'obtenir directement les objets liés. Inversement, si un objet est dépendant d'un autre, la commande "belong_to" sera utilisée, créant ici une méthode offrant un accès au parent de l'objet en question. Ces mots-clés établissant les relations peuvent aussi être utilisés afin de lier deux classes ne partageant pas de connexion directe, en utilisant la forme "has_many, :through". Dans ce cas précis, les fonctions internes de Rails se chargeront de joindre les tables et ainsi d'obtenir l'ensemble des objets liés de façon indirecte. Par exemple, en établissant qu'une mesure appartient à une source via une station, Rails crée une méthode permettant d'accéder directement à l'objet source auquel appartient une mesure. Les méthodes créées sont unidirectionnelles, si bien qu'il est nécessaire de définir "has_many" dans le modèle de la classe parent et "belongs_to" pour la classe dépendante afin que l'accès soit possible depuis chaque objet.

Finalement, les modèles permettent de définir des validations internes à Rails qui ne modifient pas la structure de la table. Celles-ci, tel que mentionné précédemment, se font à un coût de calcul plus élevé que les vérifications au niveau du SGBDR mais peuvent être modifiées sans devoir créer des migrations et altérer la structure de la base de données. Elles sont aussi plus versatiles, et permettent aussi de définir des contraintes sur plusieurs champs qui incluraient des champs vides, ce qui n'est pas possible directement dans PostgreSQL. L'utilisation de la commande "validate_uniqueness_of", suivi du ou des champs, permet de définir les champs dont l'unicité doit être garantie au moment de l'insertion ou de la modification de valeurs dans la base de données.

L'objet source est défini au Code 4.8. La première ligne présente le nom de l'objet et le définit comme une extension du module de gestion de base de données de Rails. Les attributs publics de l'objet sont définis à la deuxième ligne. Dans le cas de l'objet source, ces attributs représentent intégralement les deux champs de la table définie plus tôt. Les relations sont par la suite établies, les sources possédant plusieurs stations, ainsi que plusieurs mesures indirectement via l'objet station. Rails n'établit pas de limite d'accès, et permet par conséquent, aux lignes 5 à 8, d'établir des relations avec les paramètres, les types, les unités et les lots de façon indirecte en passant par les mesures définies précédemment.

Finalement, la ligne 9 fait état d'une validation selon le nom, ce qui fait en sorte qu'un

nom ne peut être utilisé que par un seul objet source.

Code 4.8: Modèle pour l'objet Source - source.rb

```
1 class Source < ActiveRecord::Base
2   attr_accessible :nom, :note
3   has_many :stations
4   has_many :mesures, :through => :stations
5   has_many :types, :through => :mesures
6   has_many :parametres, :through => :mesures
7   has_many :unites, :through => :mesures
8   has_many :lots, :through => :mesures
9   validates_uniqueness_of :nom
10 end
```

Le modèle pour les stations est présenté au Code 4.9. La ligne 2 permet de définir l'information appropriée pour le champ géographique PostGIS. Il importe, pour ces cas, de spécifier un système de référence dans lequel les informations seront entrées. Dans le cas présent, le système utilisé est l'ellipsoïde de référence WGS84 associé l'identifiant numérique 4326, qui permet de créer des entrées géographiques avec une localisation sous la forme de latitude et de longitude. L'utilisation de ce système de référence dans le modèle permet de créer des objets géographiques grâce au format WKT [9], en utilisant une structure semblable à POINT(longitude latitude). À noter que si les objets géographiques devaient être créés dans un autre système de référence, l'extension RGEO, qui est utilisée afin de permettre l'accès aux fonctions de PostGIS, ferait en sorte que le positionnement soit ramené au WGS84 défini dans le fichier modèle. Le stockage d'un objet géographique PostGIS se fait uniquement dans le champ position, les champs latitude et longitude créés précédemment étant ajoutés afin de permettre à quelqu'un qui ferait l'observation de la table de la base de données de pouvoir lire directement les informations de localisations.

La ligne 3 utilise la notation "belongs_to" pour définir la relation d'appartenance de la table avec les sources, alors que les lignes 5 à 8 définissent l'ensemble des tables liées directement ou indirectement aux stations. Comme pour les sources, l'ensemble des attributs d'un objet station qui sont accessibles directement doivent être définis individuellement, ce qui est fait à la ligne 4.

Les relations internes entre des stations parents et des sous-stations sont définies aux lignes 9 et 10. Tel que mentionné précédemment lors de l'élaboration du schéma de données, une

9. Well Known Text : Un format de délimitation de texte

station peut appartenir à une station “parent” (ligne 9) en plus de posséder des sous-stations “enfants” (ligne 10).

Finalement, des validations d’unicité de l’objet entré sont établies à la ligne 11, spécifiant qu’une seule combinaison de code, de source, de direction, de voie et de parent peut exister pour un objet station (la sous-section 4.2.2 fait état de la justification de cette structure).

Code 4.9: Modèle pour l’objet Station - station.rb

```
1  class Station < ActiveRecord::Base
2    set_rgeo_factory_for_column(:position, RGeo::Geographic.spherical_factory(:srid
         => 4326))
3    belongs_to :source
4    attr_accessible :alias, :nom, :nom_court, :position, :source, :source_id, :code,
         :parent_id, :direction, :voie, :parent, :latitude, :longitude
5    has_many :mesures
6    has_many :lots, :through => :mesures
7    has_many :parametres, :through => :mesures
8    has_many :types, :through => :mesures
9    belongs_to :parent, :class_name => 'Station'
10   has_many :children, :class_name => 'Station', :foreign_key => 'parent_id'
11   validates_uniqueness_of :code, :scope => [:source_id, :direction, :voie,
         :parent_id]
12 end
```

L’objet type se révèle plus simple, et est présenté au Code 4.10. L’objet étant indépendant, la définition de ses attributs accessibles en ligne 2 est suivi des relations le liant aux 6 autres objets. Finalement, une validation assurant l’unicité de chaque groupe de nom, sous_type et intervalle doit être définie.

Code 4.10: Modèle pour l'objet Type - type.rb

```
1 class Type < ActiveRecord::Base
2   attr_accessible :intervalle, :nom, :sous_type
3   has_many :mesures
4   has_many :parametres, :through => :mesures
5   has_many :unites, :through => :mesures
6   has_many :stations, :through => :mesures
7   has_many :lots, :through => :mesures
8   has_many :sources, :through => :stations
9   validates_uniqueness_of :nom, :scope => [:sous_type, :intervalle]
10 end
```

Comme pour l'objet type, les objets paramètre et unité sont assez simples et sont définis respectivement au Code 4.11 et au 4.12. Dans les deux cas, tous les attributs sont rendus disponibles, alors que les relations autant directes qu'indirectes sont définies subséquemment. Les paramètres présentent une contrainte d'unicité sur trois champs, soit le nom, la vitesse minimale et la vitesse maximale, tel que stipulé précédemment, alors que seule une vérification sur le nom est faite dans le cas des unités.

Code 4.11: Modèle pour l'objet Paramètre - parametre.rb

```
1 class Parametre < ActiveRecord::Base
2   attr_accessible :nom, :nom_html, :valeur_max, :valeur_min
3   has_many :mesures
4   has_many :stations, :through => :mesures
5   has_many :types, :through => :mesures
6   has_many :unites, :through => :mesures
7   has_many :lots, :through => :mesures
8   has_many :sources, :through => :stations
9   validates_uniqueness_of :nom, :scope => [:valeur_max, :valeur_min]
10 end
```

Code 4.12: Modèle pour l'objet Unité - unite.rb

```
1 class Unite < ActiveRecord::Base
2   attr_accessible :multiple, :nom, :nom_html
3   has_many :mesures
4   has_many :parametres, :through => :mesures
5   has_many :stations, :through => :mesures
6   has_many :types, :through => :mesures
7   has_many :lots, :through => :mesures
8   has_many :sources, :through => :stations
9   validates_uniqueness_of :nom
10 end
```

Le modèle pour l'objet lot est défini assez simplement au Code 4.13, la plupart des contraintes étant définies dans la migration créant la table. Tous les liens doivent à nouveau être définis, alors qu'une validation vise à éviter que les mêmes fichiers d'origine soient insérés à plusieurs reprises dans le système d'information.

Code 4.13: Modèle pour l'objet Lot - lot.rb

```
1 class Lot < ActiveRecord::Base
2   has_many :mesures
3   has_many :stations, :through => :mesures
4   has_many :types, :through => :mesures
5   has_many :parametres, :through => :mesures
6   has_many :unites, :through => :mesures
7   has_many :sources, :through => :stations
8   attr_accessible :fichiers_source, :horodatage, :usager
9   validates_uniqueness_of :fichiers_source
10 end
```

La structure du modèle pour les mesures est à nouveau très semblable à ceux présentés précédemment et est illustrée au Code 4.14. Si la relation de dépendance envers les autres objets est présentée aux lignes 2 à 6, la ligne 7 présente un cas unique. La table mesure est en effet la seule présentant une relation d'appartenance de deuxième niveau. En effet, toutes les autres relations indirectes se font en lien avec un intermédiaire qui fait en sorte que chacun des objets concernés peut "posséder" plusieurs objets de l'autre type (relation "many-to-many"). Dans ce cas, une source peut présenter plusieurs types de données, et chaque type peut provenir de plusieurs sources. Toutefois, si une source présente un nombre aussi grand

que désiré de mesures, une mesure ne peut être liée qu'à une seule source. Afin d'exprimer la relation, Rails ne permet pas d'utiliser une notation semblable à "belongs_to", et il est nécessaire d'utiliser la notation "delegate" en spécifiant l'intermédiaire, tel que stipulé à la ligne 7.

Finalement, les contraintes associées à l'objet mesure étant définies au niveau de la base de données pour des questions de performance du système, il n'est pas nécessaire d'inclure des validations au niveau du modèle pour la classe Mesure, celles-ci étant intégrées dans le fichier de migration présenté au Code 4.7.

Code 4.14: Modèle pour l'objet Mesure - mesure.rb

```
1 class Mesure < ActiveRecord::Base
2   belongs_to :station
3   belongs_to :parametre
4   belongs_to :type
5   belongs_to :unite
6   belongs_to :lot
7   delegate :source, :to => :station
8   attr_accessible :horodatage, :qualite, :valeur, :station, :parametre, :unite,
        :type, :lot, :station_id, :parametre_id, :type_id, :unite_id, :lot_id
9 end
```

4.5 Nature itérative du développement

La construction du schéma, l'analyse des objets et le montage du système d'information sont ici montrés comme un processus linéaire dans le but de simplifier l'explication de la démarche entreprise. Dans les faits, le développement du schéma s'est fait de manière itérative dans une approche semblable à la méthode de développement Agile.

Le déroulement du processus s'est fait en fonction de la mise en disponibilité des ensembles de données, soit dans un ordre semblable à celui utilisé au Chapitre 3. Par conséquent, les objets ont évolué au fur et à mesure que de nouveaux ensembles de données étaient étudiés et que le processus d'incorporation de ceux-ci dans le système d'information était entrepris. Par exemple, dans la première étape de développement, alors que seulement les données de qualité de l'air étaient étudiées, la notion de hiérarchie des stations était totalement inexistante. Dans le même ordre d'idée, certaines des contraintes et relations ne sont apparues que lorsque les structures changeantes forçaient une réorganisation partielle ou totale de la structure des objets. Notamment, les contraintes pour la classe Paramètre ont été modifiées suite à l'ajout de champs permettant de stipuler des catégories étant définies par des intervalles.

D'autres objets étaient totalement inexistants lors des premières étapes. C'est notamment le cas de l'objet lot, qui n'est apparu que lorsque des problèmes de qualité de données ont été identifiés et ont engendré le besoin de revenir sur une insertion de données. Le besoin de reprendre totalement une telle opération, qui passait alors par une remise à zéro de la base de données et à la réinsertion de l'ensemble des informations, a mené à la création de champs permettant de conserver des traces de l'insertion afin de pouvoir éliminer seulement un ensemble d'informations liées à un fichier présentant des défaillances.

La nature évolutive des analyses et des formats de données nécessaires afin de les réaliser fait en sorte que la méthode de développement Agile se révèle particulièrement efficace, celle-ci permettant la construction d'un système d'information s'adaptant facilement et rapidement lors de l'apparition de nouveaux besoins.

4.6 Synthèse

Le présent chapitre a permis de définir une structure qui sera commune à l'ensemble des données étudiées au Chapitre 3. La définition d'un schéma décomposé permet de diviser chacune des lectures en un ensemble de sept objets simples et de mettre ces objets en relations les uns avec les autres. Cette décomposition des données offre plusieurs avantages, notamment d'assurer la qualité des données stockées et de contrôler la redondance de certaines informations. Elle devrait aussi permettre d'insérer de nouveaux ensembles de données si nécessaire sans nécessiter de changements majeurs qui forceraient à modifier les mesures déjà accumulées dans le système.

Suite à la définition de cette structure, plusieurs choix technologiques ont été faits, permettant de faire l'implantation du schéma. Les principaux logiciels choisis, soit Ruby on Rails et le couple PostgreSQL et PostGIS, permettent de faire une implantation facile du schéma et des contraintes ainsi que de peupler le système d'information avec les données disponibles, étape qui sera présentée au chapitre suivant.

CHAPITRE 5

TRAITEMENTS AUTOMATISÉS, POTENTIALITÉS ET ANALYSES

Ce chapitre a pour principal objectif de relater les travaux permettant de faire le peuplement de la base de données, de détailler certains traitements et analyses automatisés, ainsi que d'analyser certaines capacités du système d'information développé. Un ensemble d'opérations automatisées, notamment des opérations d'insertion de données, de production de tableaux et graphiques et d'analyses multivariées seront d'abord abordées. Une deuxième section portant sur les capacités du système se penchera quant à elle sur la performance générale du système ainsi que sur les possibilités d'extension existantes qui devraient permettre d'assurer le soutien à l'analyse pour les étapes subséquentes du mandat de recherche.

5.1 Opérations automatisées

Comme mentionné à plusieurs reprises, la taille des données fournies fait en sorte que des traitements automatisés sont nécessaires dans le but de faire l'insertion des données dans le système d'information. De plus, des disparités entre les données, notamment au niveau des intervalles entre les mesures, font en sorte que des travaux subséquents peuvent être requis afin de pouvoir avoir accès à l'ensemble des informations à un niveau de résolution précis tout en pouvant stocker les informations de la façon la plus détaillée possible. Finalement, les outils permettant d'avoir des accès aux données via un langage de programmation de haut niveau ainsi que le besoin de répéter certaines analyses sur plusieurs années dans le cadre d'un mandat de recherche font en sorte qu'il peut être intéressant de définir des opérations automatisées pour exécuter ces travaux.

5.1.1 Insertion des données

Le processus d'insertion des différents types de données est de façon générale assez similaire pour chacun des types de fichiers identifiés, une procédure séquentielle générale étant appliquée dans tous les cas. Tout d'abord, dans le cas des formats de données qui se présentent sous une forme autre que des fichiers texte, ceux-ci doivent être convertis vers un format CSV, qui est beaucoup plus facile à interpréter dans Ruby via la librairie intégrée. S'il est assez simple de procéder à de telles conversions et d'obtenir un standard simple pour les données de qualité de l'air vu le nombre limité de formats, il est nécessaire de procéder de façon manuelle avec les ensembles de données de circulation.

Suite à l'importation des fichiers texte à l'intérieur d'un processus Rails, la plupart des scripts menant à l'insertion des données prennent une forme semblable, qui peut être décrite par l'algorithme suivant :

1-Interprétation du fichier (lecture automatisée du fichier selon le format attendu, assignation des valeurs à des variables)

2-Pour tout fichier, identifier des paramètres, stations, unités, types uniques

3-Pour chacun des paramètres, stations, unités, types, faire l'insertion d'un objet, en demandant à l'utilisateur de fournir des informations supplémentaires si nécessaire (obtenues des dictionnaires)

4-Lire chaque ligne, individuellement, en appliquant, si nécessaire :

4.1-Un changement d'heure

4.2-Une vérification si la mesure doit être insérée [1]

5-Procéder à une récupération en cas d'erreur causée par la présence de doublons (si une erreur est générée pour cette raison)

6-Procéder à l'insertion d'un objet lot contenant le nombre de mesures associées et le ou les fichiers d'origine

Les opérations détaillées dans la liste précédente ne sont bien sûr qu'un aperçu des procédures, et chaque type de données présente un ensemble de variations qui lui sont propres, et qui ne seront pas détaillées ici dans l'ensemble par souci d'espace. Certains extraits de code jugés plus importants seront présentés dans le présent chapitre, alors que d'autres seront fournis en Annexe.

Bases de données Access

Une première procédure d'importance consiste à procéder à l'exportation des bases de données d'un format Microsoft Access vers un format texte pouvant être interprété facilement par Ruby et Rails. Afin d'obtenir de tels fichiers CSV, l'utilitaire mdbtools [2] permet de faire une conversion directe. Les formats internes des tables étant variables, il est toutefois nécessaire de procéder à des changements supplémentaires, notamment en réorganisant ou en ajoutant des colonnes présentant des informations essentielles. L'usage d'expressions régulières permet de façon générale de faire cette réorganisation de façon automatique en quelques secondes.

Une expression régulière est une chaîne de caractères répondant à un motif. En faisant des recherches et des remplacements simples, il est donc possible d'utiliser le motif d'une ligne

1. Certaines données météo doivent être ignorées compte tenu qu'elles ne contiennent aucune information
2. Le logiciel et la documentation peuvent être obtenus à l'adresse : http://mdbtools.sourceforge.net/

complète et de le réorganiser afin qu'il prenne une structure particulière. Dans le cas des données de qualité de l'air du RSQA, les modifications viseront principalement à convertir les codes numériques des paramètres en chaînes de caractères appropriées. Pour les bases de données Access des données du MTQ, il sera plutôt nécessaire d'intégrer le code numérique de la station ainsi que le paramètre étudié se trouvant dans le nom des tables dans des champs spécifiques.

Le Code 5.1 présente quelques expressions régulières utilisées dans le but de réorganiser les tables de la base de données du MTQ. Le format permettant de définir les motifs est assez simple, les chaînes de caractères situées entre les deux premières barres obliques représentant le motif à rechercher et les expressions entre les deuxièmes et troisièmes barres obliques représentant le motif de remplacement. Les parenthèses permettent quant à elle de stocker les expressions pour les réutiliser dans les motifs de remplacement à l'aide de l'indicateur $. La première ligne permet de remplacer le séparateur numérique de centième, qui est parfois une virgule, alors que la seconde élimine des guillemets inutiles. La troisième ligne permet de réorganiser le format d'horodatage extrait à l'aide des utilitaires de mdbtools. En effet, les dates extraites prennent la forme "JJ/MM/AAAA HH:MM", ce qui est incompatible avec la façon de gérer les objets horodatages dans Ruby. Il est donc nécessaire de convertir ces horodatages vers le format interne de Ruby, soit "AAAA-MM-JJ HH:MM:SS", afin de pouvoir en faire l'utilisation.

Code 5.1: Exemple d'expressions régulières utilisées pour les données du MTQ

```
1 /(\d+)(,)(\d+)/$1.$3/
2 /"//
3 /(\d{2})\/(\d{2})\/(\d{4})\s␣(\d{2}:\d{2})/$3-$2-$1␣$4:00/
```

Pour les deux sources de données de qualité de l'air (MTQ et RSQA), un format CSV présentant les champs "station, parametre, horodatage, mesure" est attendu, si bien qu'une seule procédure d'insertion dans la base de données aura à être développée.

Format des données de circulation

Comme pour les données de qualité de l'air, les données de circulation se présentent dans des formats multiples qui doivent être traités manuellement et seront ramenés à un format commun intermédiaire qui pourra être inséré facilement dans le système d'informations. Les données répondant à ce format intermédiaire dans une base de données du logiciel Microsoft Access ont été préparées par M. Philippe Gaudette dans le cadre d'un stage. Cette base de

données contient deux tables, une présentant les informations sur les stations et les fichiers d'origine, et une autre table comprenant les données accumulées par ces stations. Les champs de la table "données" incluent notamment les horodatages, la durée de l'intervalle de mesure, ainsi que les comptages de vitesses par types de véhicules. Contrairement à la hiérarchie de station utilisée dans le système d'information, le format intermédiaire associe toujours les informations de voies aux mesures plutôt qu'à une station, ce qui devra être traité lors du processus d'insertion.

Horodatages

La gestion des changements de fuseau horaire est un problème existant pour deux ensembles de données majeurs, soit les données de météo et de qualité de l'air. Dans le cas des données de qualité de l'air, ce problème s'accompagne aussi par l'absence de d'identification des données manquantes. Par conséquent, il est intéressant de développer une procédure unique afin de générer les informations quant aux changements d'heure et aux entrées vides qui doivent être prises en compte.

L'identification des moments précis de changement d'heure n'est pas une fonction qui est supportée par Ruby ni la plupart des autres langages de programmation. Malgré tout, une méthode permet d'identifier si un horodatage est à l'heure avancée ou à l'heure normale. Il suffit donc de générer tous les horodatages entre les premières et dernières heures disponibles pour un ensemble de données, et d'y identifier les points de changements d'heure. Utiliser cette procédure permet de générer tous les horodatages existants pour la période pour laquelle les mesures sont fournies, des informations qui sont utiles dans le but d'insérer des valeurs nulles pour les heures où aucune mesure n'est disponible.

La fonction définie renvoie trois informations primordiales, soit l'heure à insérer dans la base de données, une chaîne de recherche permettant de sélectionner l'entrée appropriée dans le fichier source, ainsi qu'une information si cette heure particulière est à l'heure normale ou avancée. Concrètement, la fonction, pour un horodatage extrait du fichier source présentant la valeur "2010-06-01 01:00:00", fournira un horodatage pour l'insertion dans la base de données de "2010-06-01 01:59:59", ce qui répond à la norme établie plus tôt. La fonction utilisée pour les processus d'insertion des données de qualité de l'air et de météo est fournie en Annexe A.

Entrées doubles

Le système de contraintes détaillé à la section Section 4.3.2 et implanté à la Section 4.4.2 fait en sorte que le système renvoie une erreur à chaque fois qu'une procédure tente d'insérer une entrée alors qu'une mesure existe déjà pour une combinaison de station, paramètre et

horodatage. Afin d'éviter la fin de l'exécution à chaque fois qu'une telle erreur est générée, il faut spécifier au système une procédure de secours offrant une autre option d'exécution en cas de conflit. Dans ce cas particulier, il est pertinent de conserver les entrées doubles séparément dans le but de pouvoir plus tard enquêter sur les entrées où de tels doublons existent ainsi que sur leurs fichiers d'origine.

Tous les ensembles de données étant à risque de rencontrer des entrées doubles, la procédure devra être mise en place dans l'ensemble des procédures d'insertion. Tel que décrit à la Section 3.5.1, certains cas de doublons pour une même heure surviennent à quelques reprises dans le cas des données de circulation sur les ponts. Il est donc nécessaire de définir un cas où, à l'intérieur d'un même lot, les deux données de mesures pour la même combinaison de station et paramètre doivent être additionnées.

Le Code 5.2 présente la procédure de secours dans le cas des données de circulation pour les ponts. La première ligne représente la création d'un objet mesure où les différentes propriétés de l'objet sont assignées par des variables définies précédemment. L'ensemble du bloc contenu entre la ligne 2 et la fin du code présente la tentative de sauvegarde de cet objet et la méthode de secours si jamais un conflit est détecté.

Dans le cas des données de circulation sur les ponts, la ligne 5 tente d'identifier si une entrée conflictuelle provenant du même lot cause le problème d'unicité, et le cas échéant, additionne à cette entrée conflictuelle la valeur en cours avant de la sauvegarder (lignes 9 et 10). Pour les autres cas, c'est-à-dire lorsqu'une entrée déjà présente dans le système d'information entre en conflit avec la tentative d'insertion en cours et ne fait pas partie du même lot, la ligne 7 ajoute les informations de l'objet mesure en cours à une variable qui stocke sous format CSV l'ensemble des doublons pour permettre à l'utilisateur d'effectuer des vérifications ultérieures.

Code 5.2: Gestion des entrées doubles

```
1 m = Mesure.new(:station => s, :parametre => p, :horodatage => horodatage, :valeur
    => i[:nombre], :type => t, :unite => u, :lot => l)
2 begin
3     m.save
4     rescue ActiveRecord::RecordNotUnique
5         entree_conflictuelle_lot = Mesure.find_by_station_id_and_parametre_id
              _and_horodatage_and_type_id_and_unite_id_and_lot_id(s, p, horodatage, t,
              u, l)
6         if entree_conflictuelle_lot.nil?
7             doublons << s.id.to_s + "," + p.id.to_s + "," +
                  horodatage.strftime(timeoutformat) + "," + i[:nombre].to_s + "," +
                  t.id.to_s + "," + u.id.to_s + "\n"
8         else
9             entree_conflictuelle.valeur += i[:nombre].to_i
10            entree_conflictuelle.save
11        end
12 end
```

Les entrées conflictuelles étant stockées dans un format CSV qui contient les identifiants des autres objets, il est aussi possible de définir une procédure permettant d'effectuer la mise à jour des données déjà stockées dans le système d'information à partir de ce fichier. Une telle procédure pourrait s'avérer utile dans le cas où des données fournies feraient l'objet de révision ou si la période de nouveaux ensembles chevauchait celle de données déjà intégrées. Dans de tels cas, les nouvelles entrées pourraient être ajoutées directement à la base de données dans un processus d'insertion classique, et les entrées conflictuelles pourraient par la suite être mises à jour indépendamment.

Intervalles de mesure pour les données de circulation

Les données de comptage sur les ponts présentent une autre problématique, soit d'avoir plusieurs intervalles de mesures à l'intérieur d'un même fichier. Puisqu'un cas de chevauchement entre des données à intervalle d'une heure et à intervalle de quinze minutes existe, il n'est pas prudent de procéder de façon totalement automatisée à la gestion de ce problème.

Certaines situations récurrentes dans les fichiers permettent toutefois de définir des procédures semi-automatisées. Tout d'abord, les changements d'intervalles de mesures ont la très forte tendance à se faire lorsque des pannes surviennent. Par conséquent, tous les sauts dans le temps de plus d'une heure devraient être identifiés, et une vérification faite sur les données

avant et après afin d'essayer d'y déterminer les intervalles. Dans ces cas précis, extraire les données entre les sauts temporels dus à des données manquantes et y faire la moyenne des intervalles entre les horodatages devrait fournir une information approximative des intervalles qui y existent et permettre à l'utilisateur d'y diviser le fichier s'il y a lieu. La moyenne permet de faire une approximation de l'intervalle et d'utiliser cette information lors de l'insertion des données.

Les lectures présentant des intervalles de mesures de quinze minutes à l'intérieur de groupe dont les mesures ont supposément des intervalles d'une heure peuvent aussi être identifiées facilement en vérifiant que les chaînes de caractères ne comprennent pas les valeurs "15", "30" et "45". L'évaluation du cas inverse est toutefois impossible et devra faire l'objet de traitements manuels. La procédure doit donc identifier tous les sauts dans le temps de plus d'une heure, faire des moyennes sur les périodes entre ces sauts, et afficher l'information afin que l'utilisateur puisse investiguer et séparer les fichiers aux besoins en fonctions des intervalles. La séparation des fichiers est nécessaire en fonction des intervalles est nécessaires afin d'associer la bonne valeur d'intervalle à chacune des mesures.

Si la méthode peut sembler primitive, elle s'est révélée efficace afin d'identifier les discordances présentes dans les fichiers. Dans les ensembles de données fournis, seul le cas unique de chevauchement mentionné à la Section 3.5.1 n'a pu être identifié. Dans ce cas particulier, les contraintes d'unicité ont mis fin à l'exécution du processus d'insertion et l'utilisation de la méthode de secours présentée à la Section 5.1.1 et l'investigation subséquente ont rendu possible l'identification du problème.

Données météo

Le format des données météo diffère assez fortement des autres ensembles de données qui prennent la forme de fichiers CSV une fois transformés automatiquement ou manuellement. En effet, les fichiers de données météorologiques présentent un ensemble de 24 mesures par ligne représentant une journée complète de lectures. Puisque cette structure est standardisée et que de façon générale, ce standard est respecté, il est possible de définir un algorithme permettant de transformer les fichiers texte en un format utilisable de façon automatisée.

La meilleure façon de procéder afin de convertir les données en un format utilisable est d'exploiter le fait que Ruby traite les chaînes de caractères comme des tableaux. Puisque les délimitations sont bien définies dans le format, il suffit alors de lire le fichier ligne par ligne, et d'assigner à des variables les informations pertinentes. Le code numérique de la station peut être déterminé en faisant l'extraction des caractères 0 à 6, alors que le code numérique du paramètre mesuré se présente aux positions 19 à 23. En définissant les positions des mesures et en exécutant une boucle sur l'ensemble de celles-ci, l'horodatage complet peut

être reconstitué et la mesure récupérée. Finalement, l'ensemble de ces informations peut être rassemblé dans un tableau associatif unique, qui pourra être utilisé d'une façon semblable aux données importées de qualité de l'air ou de circulation. La fonction permettant d'exécuter les importations de ces données météo peut être étudiée à l'Annexe B.

Résultats de l'insertion

L'exécution des processus d'insertion pour toutes les données obtenues dans cette première phase a pour résultat de peupler les tables représentant les différents objets énoncés au Chapitre 4. Le Tableau 5.1 présente le nombre d'entrées pour chacun des objets. Un autre tableau faisant état de la structure de la table station, et incluant le nombre de mesures associées à chacune d'entre elles est disponible à l'Annexe C.

Tableau 5.1: Objets créés par le processus d'insertion

Objets	Compte
Source	3
Station	240
Paramètre	412 [3]
Type	12
Unité	8
Lot	54
Mesure	12 732 093

Les temps d'exécution sont très variables selon les versions du SGBDR et selon la puissance de l'ordinateur utilisé. L'insertion des données de qualité de l'air du MTQ et du RSQA prennent chacun en moyenne quatre à cinq heures, les données de météo pour une année un peu plus de dix heures, alors que les données de circulation sur les ponts peuvent prendre plus de 40 heures. Les processus d'insertion ont été exécutés sur plusieurs ordinateurs personnels utilisant des systèmes d'exploitation divers, avec des temps d'exécution assez constants, peu importe le matériel utilisé. Des gains de performance mineurs ont tout de même pu être identifiés sur des ordinateurs utilisant des stockages électroniques plutôt que des disques durs magnétiques classiques, mais les bénéfices associés n'ont pas été mesurés formellement. Dans tous les cas, il y aurait probablement lieu d'effectuer des optimisations qui viendraient réduire

3. Cette valeur ne tient pas compte des équivalences de COV développées afin de réconcilier les nomenclatures anglophones et francophones.

les temps d'exécution, mais de tels travaux n'ont pas été jugés assez importants pour être faits immédiatement.

Un total de six scripts permettant de faire l'insertion ont été développés. Un de ceux-ci fait l'intégration des données de qualité de l'air à partir du format intermédiaire énoncé à la Section 5.1.1. Deux scripts séparés sont utilisés pour les données de COV, les structures fondamentales des fichiers fournis diffèrent fortement. Les données de circulation sont traitées de la même façon, un script faisant l'insertion des données pour les ponts alors qu'un autre prend en charge le format intermédiaire développé plus tôt à la Section 5.1.1. Finalement, les données météo sont traitées par un seul programme permettant d'interpréter et de faire l'insertion du format texte fourni en utilisant l'algorithme présenté à la Section 5.1.1.

5.1.2 Niveaux de résolution

Un des problèmes majeurs rencontrés dans la gestion des données provenant de sources multiples se présente sous la forme de standards de mesure variables selon les exploitants ou selon les périodes d'échantillonnage. Dans les données disponibles dans le cadre de ce projet, ce phénomène est particulièrement présent dans le cas des données de circulation pour les ponts, qui passent fréquemment d'intervalles de mesures de 15 minutes à des intervalles d'une heure. Or, en fonction de ce qui doit être analysé, les deux niveaux de résolution peuvent se révéler pertinents. Il importe alors de procéder au stockage de ces informations d'une façon qui permettra d'afficher les données selon les deux échelles temporelles en fonction des besoins.

Le langage SQL présente une façon de réconcilier ces besoins avec l'aide d'une vue, soit une table virtuelle créée suite à une requête sur une table réelle. Dans les faits, cette table virtuelle peut alors présenter le même ensemble de données sous une forme agrégée, dans le cas actuel pour des intervalles de mesure d'une heure. Le Code 5.3 fait état de la requête permettant de reconstituer une table affichant l'ensemble des données de circulation insérées sur une base horaire à partir de données à une résolution temporelle plus fine.

Code 5.3: Agrégation des données de circulation sur une heure

```
1 CREATE OR REPLACE VIEW private.vue_circulation_horaire AS
2 (SELECT station_id,
3             horodatage::date AS date,
4             to_char(horodatage, 'YYYY-MM-DD HH24:59:59')::timestamp AS horodatage,
5             sum(valeur) AS somme
6      FROM private.mesures
7      WHERE type_id IN
8              (SELECT id
9               FROM private.types
10              WHERE nom = 'Circulation'
11                  AND sous_type != 'Vitesse')
12     GROUP BY station_id, date, horodatage
13     ORDER BY station_id, date, horodatage);
```

La requête fait en sorte d'assembler les données sur plusieurs critères. Puisque l'on désire faire l'addition de toutes les mesures pour chacune des heures de toutes les journées pour toutes les données de circulation, à l'exception de celles représentant des mesures de vitesses, il est nécessaire de procéder à l'extraction de tous les éléments constituant l'horodatage. Les lignes 3 et 4 du Code 5.3 permettent d'extraire respectivement la date de la prise de mesure, ainsi que l'horodatage de fin de l'heure pour toutes les valeurs contenues à l'intérieur d'une heure particulière. La ligne 4 est la plus importante, car elle permet de définir pour chaque entrée l'horodatage de fin de l'heure, qui est utilisé afin de regrouper les valeurs. PostgreSQL ne permettant pas aisément de définir un horodatage de fin de l'heure, il est nécessaire de le définir formellement en format texte ("to_char(horodatage, 'YYYY-MM-DD HH24:59:59')"), avant de convertir ce texte en un objet horodatage. La ligne 2 permet de conserver les informations quant à la station d'origine, qui seront nécessaires lors de l'étape suivante où les données devront être regroupées en fonction des stations directionnelles afin de connaître les flots de circulation pour une artère routière complète. Les paramètres variables en fonction des classes de véhicules et de vitesses sont naturellement intégrées par la requête, bien qu'il serait possible de faire d'autres regroupements pour tenir compte de ces informations. Suite à l'agrégation des données à l'heure, 855 463 entrées sont comptabilisées, comparativement à 11 840 421 pour la table mesures.

S'il peut être intéressant de faire la somme sur les voies pour chaque heure, et ainsi obtenir des mesures horaires, l'utilisation de données pour une direction entière reste la plus pertinente dans le but de faire des analyses à grande échelle. Il est possible de définir une requête permettant la création d'une vue qui maintiendra les données à jour à ce niveau de

résolution spatiale et temporel en réutilisant la vue précédente. La requête permettant de créer cette vue est présentée au Code 5.4.

Code 5.4: Agrégation des données de circulation sur une heure (par direction)

```
1  CREATE OR REPLACE VIEW private.vue_circulation_horaire_directionnel AS
2  SELECT station AS station_id,
3          (SELECT id FROM private.parametres WHERE nom = 'Indetermine' AND
               valeur_max IS NULL AND valeur_min IS NULL) AS parametre_id,
4          (SELECT id FROM private.types WHERE nom = 'Circulation' AND sous_type =
               'Comptage' AND intervalle = 3600) AS type_id,
5          (SELECT id FROM private.unites WHERE nom = 'Vehicules') AS unite_id,
6          horodatage AS horodatage,
7          somme_dir AS valeur
8     FROM (SELECT
9          (SELECT parent_id FROM private.stations WHERE id=station_id and station_id
               != parent_id)AS station,
10         (SELECT direction FROM private.stations WHERE id=station_id)AS direction,
11         horodatage::timestamp,
12         sum(somme) as somme_dir
13       FROM private.vue_circulation_horaire
14       GROUP BY station, direction, horodatage
15       ORDER BY station,horodatage) as sous_requete;
```

La requête du Code 5.4 reprend l'ensemble des mesures de circulation horaire rassemblées précédemment au Code 5.3, mais en fait cette fois l'agrégation selon la station parent. L'ensemble des stations utilisant le même système de classement hiérarchique, les mesures pour chacune des voies seront additionnées pour être liées à la station représentant l'ensemble des voies pour cette direction. Dans les cas où les stations ne présentent pas de direction ou encore une seule voie, la valeur de "parent_id" sera identique à "station_id", et ces données seront exclues. Cette situation s'explique par la volonté d'intégrer ces données agrégées au reste de la table mesures. Si celles-ci étaient regroupées, il en découlerait l'apparition de dédoublement lors de la fusion de la vue avec la table mesures d'origine. La ligne 14 est celle effectuant le regroupement des valeurs en fonction des stations parent et des directions avec la clause "group by". L'agrégation par lien routier permet de réduire le nombre de mesures à 340 745, et de faire passer le nombre de stations de 182 à 48.

Afin de préserver la compatibilité avec la table "mesures" d'origine, des liens avec les paramètres, unités et types appropriés sont faits. Cette compatibilité avec la table d'origine présente l'avantage de pouvoir faire l'union avec le reste des autres données et ainsi avoir

accès à une table présentant l'ensemble des mesures ramenées à une base horaire. Le Code 5.5 présente de quelle façon ce lien peut être complété sous la forme d'une vue supplémentaire.

Code 5.5: Union des mesures de circulation agrégées avec le reste des mesures

```
1 CREATE OR REPLACE VIEW private.vue_mesures_horaire as (
2   SELECT station_id,
3          parametre_id,
4          horodatage,
5          valeur,
6          type_id,
7          unite_id
8    FROM private.mesures
9   UNION
10  SELECT station_id,
11         parametre_id,
12         horodatage,
13         valeur,
14         type_id,
15         unite_id
16   FROM private.vue_circulation_horaire_directionnel
17 );
```

Les lignes 2 à 8 stipulent de sélectionner l'ensemble des données de la table mesures, qui seront ajoutées grâce aux vues créées précédemment. La commande "UNION" à la ligne 9 permet quant à elle de concaténer directement des tables présentant des structures semblables, ce qui est ici fait. Il n'est pas problématique ici de joindre les données de circulation directionnelles sur une heure aux données d'origine à l'intérieur d'une même table étant donné que la structure hiérarchique fait en sorte que les mesures sont associées à des stations différentes. L'agrégation sur une base horaire et directionnelle des stations de circulation fait passer le nombre de mesures total de 12 558 275 à 1 124 123.

La création de vues effectuant le rassemblement de données à différentes échelles spatio-temporelles n'est évidemment pas une procédure simple, et celle-ci demande une planification minutieuse, et peut même demander une restructuration de l'organisation du schéma. Toutefois, les vues en question seront recréées de façon automatisée sur demande, de telle sorte qu'elles peuvent être fonctionnelles tant et aussi longtemps que la structure fondamentale des données insérées dans le système d'information ne changera pas.

L'accès aux données de cette table virtuelle est malheureusement coûteux en temps de calcul puisque de très grands ensembles de données doivent être rassemblés en fonction de

l'heure de la prise de mesure, ce qui fait en sorte que le système doit définir de nouvelles valeurs d'horodatages pour toutes les mesures à l'aide de la procédure "to_char(horodatage, 'YYYY-MM-DD HH24:59:59')::timestamp". Cette dernière procédure est la plus coûteuse, mais est requise afin de grouper l'ensemble des mesures en fonction de l'heure et d'en faire la somme. Les sous-requêtes se révèlent aussi coûteuses, avec comme résultat que l'exécution de la requête prend en moyenne environ 40 minutes pour produire la vue directionnelle.

Afin de pallier ce problème, il est possible de faire l'utilisation de vues matérialisées, soit un ensemble de fonctions qui permettent de stocker le résultat d'une vue dans une table réelle et de mettre cette dernière à jour sur demande. Une fois ces vues matérialisées en place, il est aussi possible d'établir une connexion entre les outils de Rails et celle-ci à l'aide d'un fichier de modèle, de telle sorte que la vue matérialisée peut être utilisée comme n'importe quel autre point de stockage. Les mêmes relations spécifiées plus tôt dans les autres fichiers de modèles sont ici aussi applicables, rendant l'intégration naturelle. Le Code 5.6 présente le fichier de modèle permettant de définir l'accès aux données de la vue matérialisée représentant toutes les mesures sur une base horaire et directionnelle dans le cas des données de circulation.

Code 5.6: Modèle permettant d'accéder aux données d'une vue matérialisée

```
1 class MesureHoraire < ActiveRecord::Base
2   set_table_name "vm_mesures_horaire"
3   belongs_to :station
4   belongs_to :parametre
5   belongs_to :type
6   belongs_to :unite
7   attr_accessible :valeur, :station, :station_id, :horodatage, :direction,
        :parametre, :parametre_id, :type, :type_id, :unite, :unite_id
8 end
```

Il est important de mentionner que les diverses manipulations faites dans la création des vues et vues matérialisées ne conservent pas les index créés précédemment au Chapitre 4. Or, comme le démontrera la Section 5.2.3, des pertes de performances substantielles sont liées à l'absence de ces index. Par conséquent, toute utilisation de la vue matérialisée de mesures devrait être précédée par une réintroduction d'index prenant la même forme que ceux utilisés pour la table "mesures" d'origine. Le Code 5.7 relate les commandes nécessaires à la création de la vue matérialisée et la création des index.

Code 5.7: Commandes créant une vue matérialisée et les index appropriés

```
1 SELECT private.create_matview('private.vm_mesures_horaire',
      'private.vue_mesures_horaire');
2 CREATE INDEX station_id_idx ON private.vm_mesures_horaire(station_id);
3 CREATE INDEX parametre_id_idx ON private.vm_mesures_horaire(parametre_id);
4 CREATE INDEX type_id_idx ON private.vm_mesures_horaire(type_id);
5 CREATE INDEX horodatage_idx ON private.vm_mesures_horaire(horodatage);
```

5.1.3 Tableaux et graphiques

Si les opérations automatisées présentées dans les sections précédentes permettent de peupler la base de données et d'avoir accès aux données à différents niveaux de résolution, il est aussi intéressant d'automatiser d'autres opérations qui devront être reproduites à de multiples reprises au cours du mandat d'analyse. Parmi ces opérations, la production de tableaux statistiques de qualité de l'air ou des graphiques présentant les flots de circulation sur différentes artères routières devront être produits de façon annuelle dans le cadre du suivi. Les données ainsi que les tableaux qui en résultent ne pouvant être rendus publics au moment d'écrire ce mémoire, les deux sous-sections suivantes se concentreront sur les procédures à accomplir afin d'obtenir des résultats satisfaisants et sur la forme que prennent ces résultats.

Tableaux de qualité de l'air

Les tableaux de statistiques portant sur les données de qualité de l'air se déclinent sous plusieurs formes. La forme principale est un tableau synthèse rassemblant des informations comme le nombre de mesures disponibles, différents centiles calculés sur une période, le maximum et la moyenne observée ainsi que le nombre de dépassements d'une norme de qualité observée. Les molécules et particules obtenues sont toutes accompagnées de normes d'exposition maximale. Ces normes peuvent être sur différents intervalles d'exposition, certains paramètres ne possédant qu'un seuil une heure, alors que d'autres ont plusieurs seuils associés. Le Tableau 5.2 fait état des différentes normes pour les molécules et particules.

Tableau 5.2: Normes d'exposition - Molécules et Particules

Paramètre	1h ($\mu g/m^3$)	8h ($\mu g/m^3$)	24h ($\mu g/m^3$)
CO	34 000	12 700	-
NO_2	414	-	207
O_3	160	125	*50 [4]
PM_{10}	-	-	*50
PM_{25}	-	-	30
PST	-	-	120
SO_2	*1 300	-	288

Les normes d'exposition dans ce tableau sont établies par différentes agences gouvernementales et peuvent varier autant sur les valeurs limites que sur les intervalles. La présence de normes sur plus d'une heure a pour conséquence qu'il est nécessaire de procéder à l'utilisation de moyennes mobiles afin de valider si une heure atteint ou dépasse la norme. Concrètement, l'application d'une moyenne mobile fait en sorte que la valeur pour un horodatage particulier devient la moyenne des valeurs sur les heures précédentes aux fins de vérification des dépassements. Par exemple, pour une norme et une moyenne mobile sur huit heures, la valeur d'un jour particulier à 07:59:59 deviendrait la moyenne des valeurs obtenues de 00:59:59 à 07:59:59. Les moyennes mobiles ne sont jugées valides que si 75 % des valeurs sur cet intervalle sont disponibles. Si cette condition ne peut être remplie, une valeur nulle est attribuée. L'utilisation d'outils automatisés effectuant la compilation des tableaux permet d'avoir une bonne versatilité et d'effectuer les changements rapides lorsque ces normes changent.

La multiplication des normes, et le besoin de faire les compilations statistiques en fonction des paramètres fait en sorte qu'un tableau doit être produit par couple de paramètre et de critère. Par conséquent, un total de 13 tableaux doivent être produits afin de présenter les statistiques et les dépassements aux normes précédentes. [5] Ces tableaux devraient prendre la forme présentée au Tableau 5.3.

Un processus assez simple peut être développé afin d'atteindre les objectifs de production. En conservant l'ensemble des informations sur les normes dans un tableau, il suffit de faire une boucle sur l'ensemble des paramètres, des normes, des stations et des heures pour une période de temps désirée pour pouvoir produire les tableaux attendus. Pour les seuils de

4. Les normes accompagnées d'une astérisque sont les seuils utilisés par le RSQA

5. Les mesures de particules fines (PM_{25}) sont disponibles selon deux procédés de collecte de données et des tableaux doivent être produits pour chacun de ces procédés.

qualité de l'air sur une heure, il suffit d'extraire la valeur appropriée pour l'heure en cours et de l'ajouter dans un tableau. Dans les cas où un seuil s'applique sur plusieurs heures, il est nécessaire d'effectuer des requêtes supplémentaires sur les heures qui précèdent l'heure en cours et de faire la moyenne de l'ensemble de ces mesures. La valeur de cette moyenne, ou une valeur nulle pour les cas où moins de 75 % des valeurs sont valides, peut alors être stockée dans un tableau, qui fera à terme l'objet des mêmes traitements statistiques que le tableau obtenu directement de la base de données pour les seuils sur une heure. Par souci de simplification de l'organisation du fichier de sortie, un format HTML est utilisé, celui-ci pouvant être ouvert dans la plupart des logiciels de traitements de texte communs.

D'autres tableaux, produisant notamment des moyennes annuelles pour chaque station et paramètre, ou encore pour les COV peuvent être produits en respectant la même structure générale que celle utilisée ici.

Tableau 5.3: Forme attendue des tableaux statistiques synthèse

Station	Nombre de résultats		Centiles - ($\mu g/m^3$)				Maximum ($\mu g/m^3$)	Moyenne ($\mu g/m^3$)	Nombre de valeur dépassant la norme	
	Nombre	% valide	50	70	90	98			Nombre	Pourcentage
N2										
N1										
C2										
C1										
S2										
S1										
29										
55										
7										

Graphiques

L'autre forme de représentation visuelle des données explorée dans le cadre de ce travail se présente sous la forme de graphiques. Deux types principaux ont été produits, soit des graphiques illustrant les dépassements des seuils de qualité de l'air présentés plus tôt ainsi que des graphiques illustrant les flux moyens sur certains liens routiers en fonction de l'heure.

La représentation des dépassements des seuils de qualité de l'air peut être générée dans un format semblable à celui utilisé pour représenter les discontinuités dans les données obtenues tel que présenté notamment à la Section 3.3. Dans les deux cas, la procédure permettant de créer de tels graphiques consiste à créer un tableau en format html à l'aide d'une boucle sur l'ensemble des stations et des jours sur une période et d'assigner à chaque cellule une couleur en fonction du nombre de valeurs ou du nombre de dépassements pour une journée. Les couleurs peuvent quant à elles être définies en générant les valeurs hexadécimales en format RGB [6] à l'aide d'une fonction. Une telle fonction peut être définie afin d'être utilisable à la fois pour les graphiques de couverture temporelle et de dépassement grâce à l'ajout d'une variable permettant de renverser les couleurs et d'ainsi associer les valeurs de vert aux éléments positifs et les valeurs s'approchant du rouge à des éléments négatifs. Le Code 5.8 reprend la fonction utilisée afin de procéder à la création des deux types de graphiques avec comme entrée un nombre à virgule entre 0 et 1 représentant soit le pourcentage de valeurs valides obtenues ou le pourcentage de dépassements pour une journée.

Dans le but de maximiser l'utilisation du code et d'avoir une seule fonction pour définir tous les graphiques, l'option de définir les cellules ayant comme valeur 0 est nécessaire pour la fonction puisse remplir son rôle pour les graphiques de couverture temporelle. Lorsque la valeur zero_noir est définie comme vraie, la ligne 2 permet d'obtenir la couleur noire ("#000000") en sortie. Les graphiques se définissant sur une échelle allant du vert au rouge, seules ces deux couleurs seront modifiées, la ligne 3 définit par conséquent la valeur de bleu comme étant 0. La première valeur étant vert si l'échelle n'est pas inversée, et une transition graduelle permettant de passer du vert au rouge de façon graduelle avec le jaune comme couleur intermédiaire étant désirée, deux sous-fonctions sont nécessaires. Pour les valeurs entre 0 et 50 %, la valeur du vert est fixée à 255, et la valeur de rouge sera graduellement augmentée, pour atteindre le maximum à mi-chemin ("#FFFF00"). Inversement, pour les valeurs entre 50 et 100 %, la valeur du vert est diminuée pour atteindre 0 lorsque la valeur de 100 % est fournie en entrée. Le résultat de cette deuxième sous-fonction retourne donc une valeur de "#FF0000", soit l'équivalent en hexadécimal de 255-0-0, qui correspond à la couleur rouge. Finalement dans le but de rendre la fonction utilisable pour les graphiques

6. Le format RGB permet de définir une couleur en spécifiant respectivement des valeurs pour les couleurs rouge, vert et bleu.

de couverture temporelle, l'échelle doit pouvoir être inversée. Il suffit alors de réassigner les couleurs dans la variable de sortie en inversant les couleurs de vert et de rouge. Les deux échelles obtenues sont présentées à la Figure 5.1.

Code 5.8: Fonction de génération de code de couleurs

```ruby
def generer_couleur(pourcentage, inverse=false, zero_noir=false)
    return "#000000" if zero_noir and pourcentage == 0
    bleu = 0.0
    if pourcentage >= 0 and pourcentage < 0.5
        vert = 1.0
        rouge = 2 * pourcentage
    elsif pourcentage >= 0.5 and pourcentage <= 1
        rouge = 1.0
        vert = 1.0 - 2 * (pourcentage-0.5)
    end
    sortie = "#%02X%02X%02X" % [rouge*255, vert*255, bleu*255]
    sortie = "#%02X%02X%02X" % [vert*255, rouge*255, bleu*255] if inverse == true
    return sortie
end
```

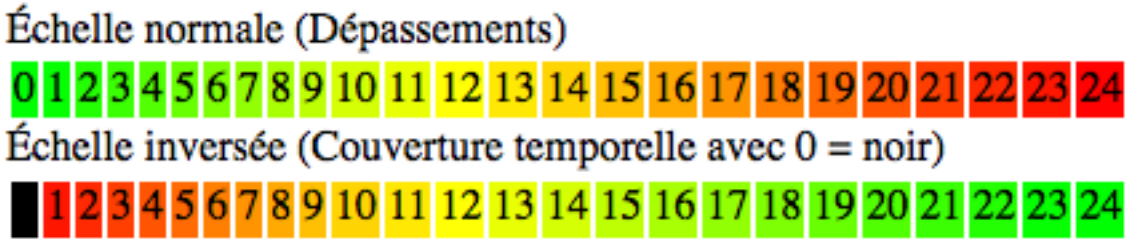


Figure 5.1: Échelles issues de la fonction définie au Code 5.8

Les graphiques qui font état des flux de circulation par direction sur les liens routiers se font simplement compte tenu de l'agrégation spatiale et temporelle réalisée à la Section 5.1.2. Une simple boucle permet d'exécuter une sélection des données pour un ensemble de liens routiers dans une direction particulière et ensuite de rassembler les données par heure pour faire une moyenne. Lorsque ces moyennes sont calculées, il ne reste qu'à produire les graphiques à l'aide d'extensions de ruby conçues pour produire de telles représentations. Un exemple des résultats qui peuvent être obtenus par cette procédure peut être observé à la

Figure 5.2. Dans ce cas particulier, la légende permettant d'identifier les liens routiers a été retirée puisque les résultats ne peuvent pour l'instant pas être publiés. L'Annexe D reproduit le code permettant de générer les graphiques.

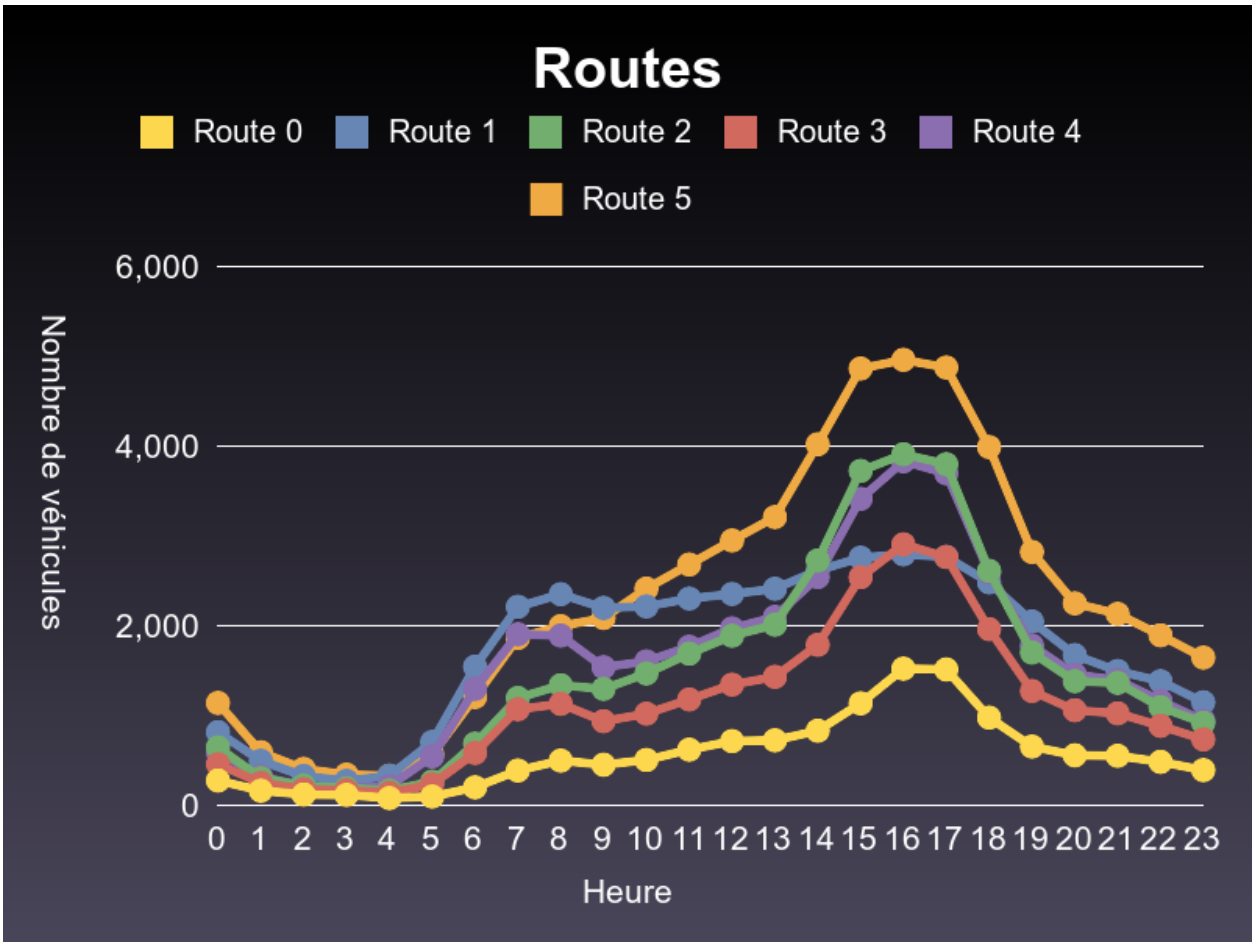


Figure 5.2: Exemple de graphique pour les données de circulation

5.1.4 Analyses multivariées

Dans certains cas, il peut être intéressant d'obtenir des valeurs en fonction de certaines variables précises. Un exemple d'une telle situation se présente lors de la mise en relation des données météo et des données de qualité de l'air. Plus précisément, il est intéressant d'obtenir les données de qualité de l'air lorsque le vent souffle perpendiculairement à l'axe de l'autoroute 25 et de comparer les résultats pour les stations des deux côtés de l'axe. Une telle analyse permet en effet de vérifier les impacts des variations météorologiques sur les concentrations de différents polluants aux stations à proximité de l'autoroute.

L'utilisation d'un système d'information intégré rend très facile de faire une requête permettant d'obtenir toutes les mesures de qualité de l'air répondant à ce critère. L'axe du pont étant dans une orientation d'environ 143°, les données d'intérêt pourraient se situer lorsque le vent souffle dans une direction de ±10°autour de la perpendiculaire à l'axe, avec comme résultat que toutes les données de qualité de l'air lorsque le vent souffle à des angles entre 43°et 63°devraient être extraites. Le Code 5.9 présente les quelques lignes de code à entrer pour obtenir l'ensemble de ces données.

Code 5.9: Sélection de données de qualité de l'air selon la direction du vent

```
1 type = Type.find_by_nom('Meteo')
2 parametre = Parametre.find_by_nom('Vent - Direction')
3 type_qa = Type.find_by_nom_and_sous_type('Qualite de l\'air', 'Molecules')
4 mesures = Mesure.where(:type_id => type_qa, :horodatage => Mesure.where(:type_id =>
      type, :parametre_id => parametre, :valeur => 43..63).pluck('horodatage'))
```

La taille de la requête est minimale dans ce cas particulier, mais elle illustre tout de même la simplicité d'imbriquer des conditions. Il est possible préalablement ou après cette opération de sélection de refaire une sélection pour obtenir un sous-ensemble concernant une station ou un paramètre précis, ainsi que d'exporter les données vers le format CSV pour en faire le traitement dans un logiciel séparé.

5.2 Potentialités

Le système d'information devant servir de répertoire central archivant l'ensemble des données liées au projet de recherche pour plusieurs années, il est intéressant d'établir sa capacité à répondre aux besoins actuels et futurs. Les capacités de connexion du système, les possibilités d'élargissement du cadre pour intégrer les données transactionnelles du nouveau pont ainsi que la capacité du système à soutenir des ensembles de données de grande taille sur plusieurs années seront principalement étudiés.

5.2.1 Accès multi-utilisateurs aux données

Un des besoins névralgiques mentionnés précédemment est la capacité du système à pouvoir servir de base de données centrale pour l'ensemble des chercheurs associés au projet de suivi des impacts du nouveau pont. Afin de servir tous ces utilisateurs, le système doit offrir la possibilité d'accéder aux données de façon distante.

Les technologies choisies permettent l'accès distant par différentes méthodes et logiciels. Il est en effet possible d'interagir avec les données stockées en procédant directement par le biais de requête SQL, soit via un logiciel offrant une interface graphique, soit par des utilitaires de ligne de commande. La plate-forme Rails peut aussi être accédée de façon distante par ligne de commande à distance, et l'accès à certaines fonctions via une interface web pourrait aussi être activé selon les besoins.

La façon la plus intéressante de se connecter à la base de données pour produire des résultats visuels est toutefois via l'utilisation des extensions PostGIS. L'utilisation de champs stockant des informations géographiques, comme le champ position de la table "stations", permet en effet de se connecter directement aux données à partir d'un logiciel de SIG. Une multitude de logiciels de cartographie possèdent les capacités de connexions à des bases de données PostGIS, les plus connus et utilisés étant le logiciel commercial ArcGIS et le logiciel libre QGIS. La Figure 5.3 présente l'écran de connexion de QGIS, en plus d'une requête simple permettant de calculer le nombre d'entrées par station pour toutes les stations ayant des mesures. La Figure 5.4 présente quant à elle le résultat obtenu dans le logiciel suite à l'exécution de la requête. L'ensemble des cartes illustrant des informations de localisation de station présentées au Chapitre 3 ont été produites à l'aide de telles requêtes.

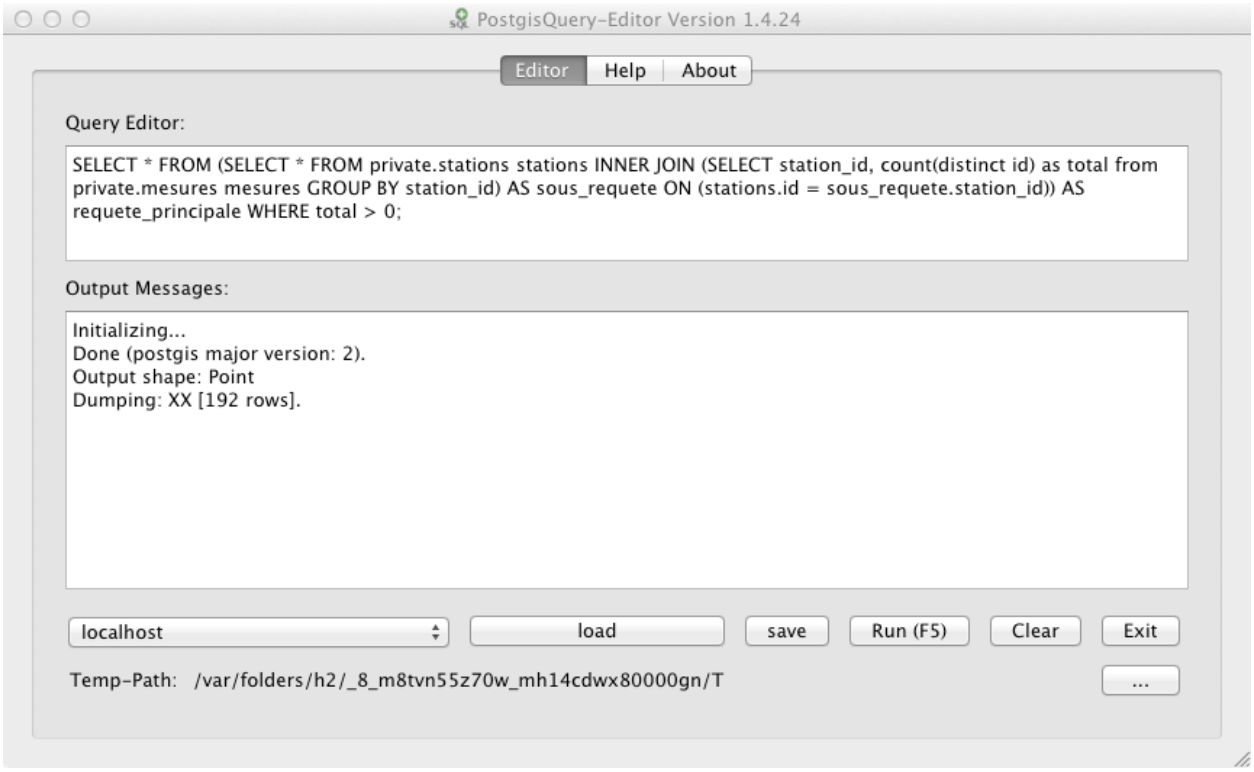


Figure 5.3: Exemple de requête PostGIS dans QGIS

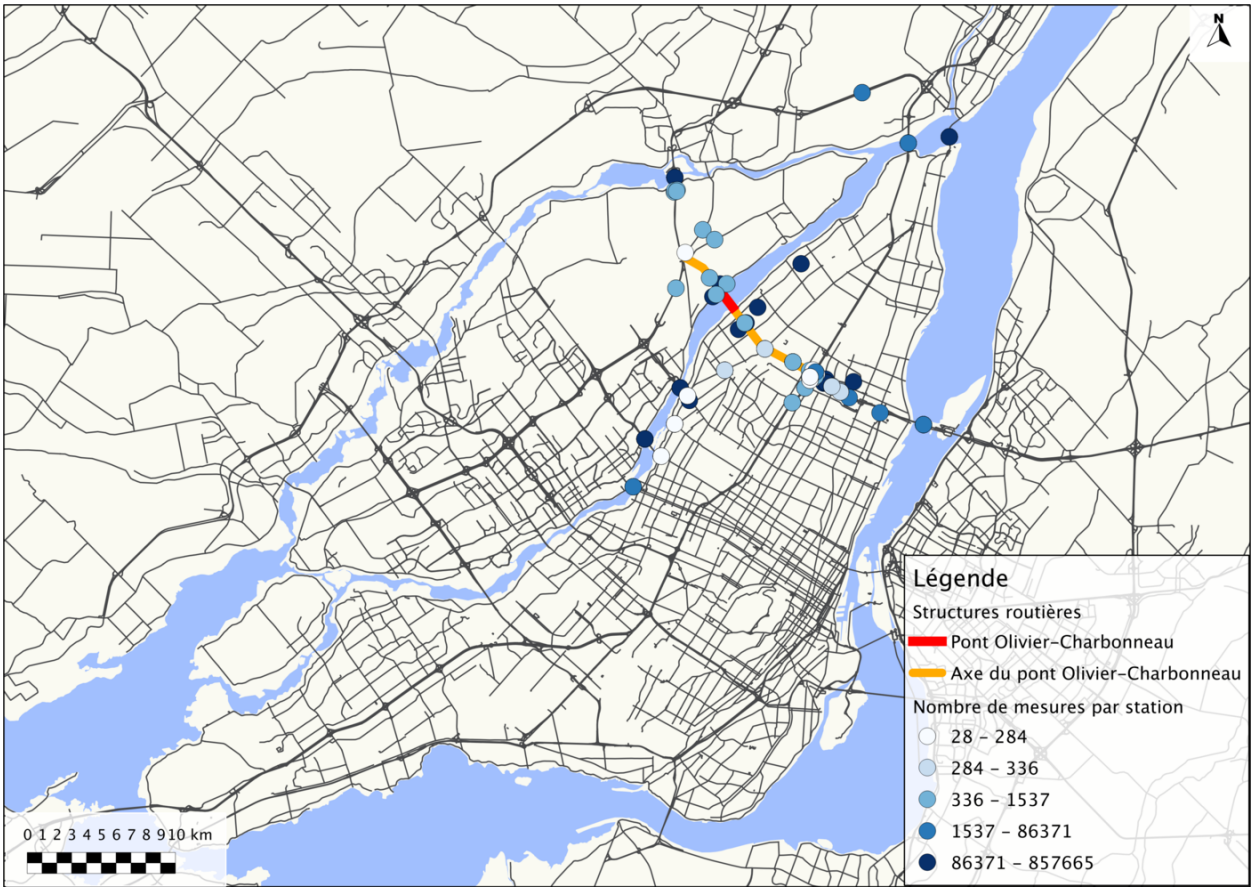


Figure 5.4: Nombre de mesures pour chaque station

La taille finale de l'ensemble des scripts développés dans le cadre de ce projet ainsi que le reste de l'application Rails est de petite taille, ce qui rend facile l'installation du système sur un ordinateur personnel. Dans le cas d'analyses plus complexes demandant de très longs temps de calcul, il est donc possible d'importer l'ensemble des fichiers, qui font moins de six méga-octets. La plate-forme Rails et les logiciels PostgreSQL et PostGIS peuvent quant à eux être installés sur la plupart des systèmes d'exploitation. Les données elles-mêmes doivent être transférées séparément et sont de plus grande taille. Une fois compressée, la base de données résultant de la première phase des travaux ne représente toutefois qu'un peu plus de 100 méga-octets, ce qui la rend toujours transférable assez facilement.

5.2.2 Élargissement du cadre d'analyse

La phase suivante du projet devrait voir l'ajout de données transactionnelles issues du système de paiement automatisé du pont. Ces données permettront une analyse plus fine

que les comptages à intervalles variables, mais devront faire l'objet de traitements afin d'être mises en commun avec les informations des autres ponts.

Le format des données n'est pas encore connu, mais il est tout de même possible d'estimer approximativement la structure qu'elles devraient prendre. En supposant que les passages de véhicules soient enregistrés de façon automatique, la forme des données pourrait être semblable à ce qui est présenté au Tableau 5.4.

Tableau 5.4: Forme anticipée des données transactionnelles

# d'identification de la transaction	# d'identification de l'abonnée ou plaque minéralogique	Horodatage	Direction	Voie ou borne d'identification de véhicule

De telles données transactionnelles seraient incompatibles avec les objets utilisés jusqu'à présent, et nécessiteraient donc la création de nouvelles tables avant de pouvoir être ramenées au paradigme utilisé pour les autres ensembles de données. Si le format des données respecte essentiellement la structure mentionnée précédemment, trois objets principaux pourraient être identifiés, soit des clients (Tableau 5.5), des véhicules (Tableau 5.6) ainsi que des transactions (Tableau 5.7). Les données devraient aussi être associées à la table stations en supposant que le matériel permettant d'identifier les véhicules dans le but d'enregistrer les transactions s'apparente à l'objet station déjà défini.

Tableau 5.5: Objet client

Champ	Type de données	Information entreposée
id	Entier auto-incrémenté	Identifiant unique
adresse	Chaîne de caractères	Adresse du client
code_postal	Chaîne de caractères	Code postal du client
position	Référence géographique	La localisation associée à l'adresse
autre	?	?

Tableau 5.6: Objet véhicule

Champ	Type de données	Information entreposée
id	Entier auto-incrémenté	Identifiant unique
plaque_mineralogique	Chaîne de caractères	Plaque minéralogique du véhicule
adresse	Chaîne de caractères	Adresse enregistrée pour le véhicule
code_postal	Chaîne de caractères	Code postal de l'adresse
position	Référence géographique	La localisation associée à l'adresse
autre	?	?

Tableau 5.7: Objet transaction

Champ	Type de données	Information entreposée
id	Entier auto-incrémenté	Identifiant unique
client_id	Entier	Référence au client
vehicule_id	Entier	Référence au véhicule
station_id	Entier	Référence à la station
horodatage	Horodatage	Date et heure du passage

En fonction des informations reçues, les objets véhicules et clients devraient intégrer des localisations, notamment sous la forme d'adresses, de code postaux ainsi qu'une géolocalisation sous la forme d'objet PostGIS. Le reste des informations seraient des champs permettant de procéder à l'identification dans le but de faire des analyses sur la clientèle utilisant le pont ainsi que ses habitudes. La structure reste toutefois approximative puisque les données ne sont pas disponibles à ce point. Par ailleurs, d'autres considérations pourraient aussi affecter la structure finale, notamment quel degré d'anonymat des clients et véhicules devrait être mis en place afin de rendre impossible l'identification formelle des transactions.

Comme dans le cas du schéma développé pour l'ensemble des données obtenues pour la première phase il serait ici aussi nécessaire d'établir des relations entre les objets définis. Dans le cas des nouvelles tables, la table "transactions" prendrait la positon centrale et dépendrait des deux nouveaux objets, soit les véhicules et les clients. Ces derniers pourraient être mis en relation en passant par la table transaction.

Finalement, l'objet transaction devrait aussi être lié à la table station, qui permet de localiser le site des transactions dans l'espace géographique. En établissant cette relation, il serait possible de ramener toutes les transactions à un format identique à celui utilisé pour

les autres données de comptage. Il serait évidemment nécessaire de créer un nouvel ensemble de stations et de sous-stations représentant le mode de collecte de données, par exemple en associant les bornes identifiant les véhicules à des voies et des directions.

Les normes hiérarchiques définies plus tôt pour les stations de circulation seraient à nouveau applicables ici, et feraient en sorte de présenter une structure semblable à celle présentée à la Figure 5.5. Dans ce système hiérarchique, les six voies numérotées représentent les points d'identification des véhicules, et peuvent être regroupées en fonction de la direction et par la suite pour l'ensemble du pont si nécessaire, comme pour les autres données de circulation.

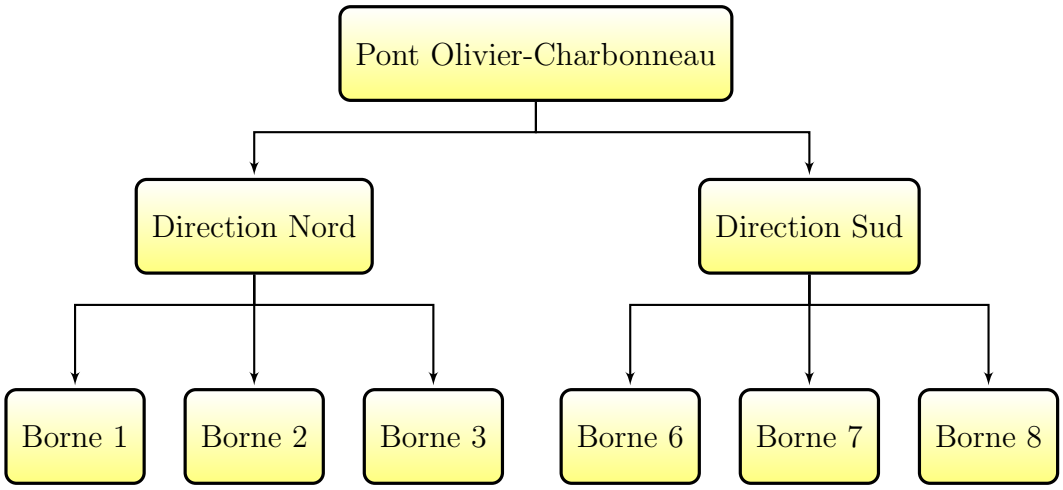


Figure 5.5: Hiérarchie des stations et sous-stations du pont Olivier-Charbonneau

Une requête unique sur la table transactions devrait alors permettre de transformer les données transactionnelles en comptages semblables à ceux déjà obtenus. En rassemblant ("group by") toutes les transactions effectuées sur une heure, les données pourraient être ramenées à une base horaire et par voies. En regroupant ces transactions sur une heure sur chaque voie en fonction des parents de chacune des sous-stations, il serait alors possible d'avoir des comptages horaires directionnels, et en amalgamant ces données à un niveau plus élevé, il serait possible d'avoir une représentation du nombre de véhicules ayant circulé sur l'ensemble des voies du pont pour une heure en particulier. Puisque la forme à ce point sera apparentée aux mesures, il ne suffira que de faire l'union entre les vues matérialisées créées à partir des données transactionnelles et les données de comptage classique contenues dans la table mesures afin de pouvoir les utiliser directement.

Comme pour les données intégrées dans la phase préliminaire, il y aurait probablement lieu de procéder à l'intégration de contraintes permettant de faire une validation des données stockées. Il est toutefois difficile d'établir formellement ces contraintes sans avoir en main les

données d'exploitation. Néanmoins, il semble acquis que des contraintes devraient être appliquées afin d'assurer l'unicité des clients ainsi que des véhicules et ainsi éviter des répétitions en plus de s'assurer que les passages sur le pont soient correctement associés.

5.2.3 Analyse de performance du système

Le système d'information développé étant appelé à servir de point d'accès principal aux données pour toutes les phases d'analyses, la quantité d'information qui y est stockée est appelée à augmenter au fil du temps. Cette augmentation devrait être nourrie par des mesures effectuées par les stations actuelles sur les années subséquentes, mais aussi sur de nouveaux ensembles de données apportant des informations complémentaires à celles déjà stockées. Dans ces circonstances, il est réaliste de penser que le nombre d'entrées dans la base de données pourrait être multiplié par trois ou quatre, ou même plus. Il est par conséquent intéressant de procéder à des analyses de performances du système afin de mesurer les coûts en termes de vitesse d'exécution associés à la charge de données entreposées.

Afin d'avoir un portrait global du comportement du système, il est important d'analyser les performances de ses principales fonctions, soit l'insertion de nouvelles données et les requêtes permettant d'avoir accès à ces données. Dans le but d'artificiellement augmenter la taille de la base de données afin de faire les tests, une procédure de clonage des données existantes a été mise en place, si bien que l'analyse se fera sur six systèmes de référence, soit un système une table "mesures" vide, une table "mesures" incluant l'ensemble des données stockées dans la première phase, et quatre autres tables ajoutant chacune une copie de cette table. Par conséquent, en définissant la table "mesures" issue de la première phase comme étant "x", le résultat sera six tables qui contiendront respectivement 0, 1, 2, 3, 4 et 5 fois le contenu des données insérées dans cette première phase. Afin d'éviter des dédoublements des horodatages, chaque fois qu'une copie du système est ajoutée, un retour en arrière de dix ans est fait, ce qui assure que les données clonées n'entreront pas en conflit avec les données déjà présentes dans le système.

Opérations d'insertion

Les opérations d'insertion sont celles qui présentent les temps d'exécution les plus longs dans les processus développés dans le cadre de ce mémoire, principalement en raison des contraintes assurant la qualité et l'organisation des données entreposées. Afin de mesurer les effets de l'augmentation du nombre de celles-ci sur les performances du système, une procédure faisant l'insertion d'un ensemble de données fictives doit être développée. Un ensemble de 25 000 données fictives, associées de façon aléatoire aux stations, paramètres, unités et

types existants est inséré par le script, ce dernier calculant aussi le temps d'exécution de l'insertion. Les temps d'exécution du processus d'insertion dans les six systèmes de référence sont présentés au Tableau 5.8, alors que le script permettant de faire l'insertion fictive est affiché à l'Annexe E.

Tableau 5.8: Performance du système lors de l'insertion de 25 000 nouvelles entrées

Taille du système	Temps d'exécution (s)	Temps d'insertion par entrée (s)
0x	223.28	0.0089
1x	266.99	0.0098
2x	276.99	0.0111
3x	323.84	0.0130
4x	346.48	0.0139
5x	377.31	0.0151

Il importe tout d'abord de mentionner que ces exécutions sont le résultat d'une situation idéalisée où aucune autre opération, telle que des changements d'horodatages, ne doit être accomplie. Il en résulte que la vitesse d'exécution est beaucoup plus rapide que pour un cas réel. À titre d'exemple, les insertions du premier ensemble de données météo, comportant neuf mois de mesures et un peu moins de 66 000 entrées, prennent en général de cinq à six heures.

Une situation idéalisée permet d'éliminer les problématiques diverses et de se concentrer sur la performance du système lui-même. Le Tableau 5.8 démontre bien la perte de performance graduelle qui survient à mesure que la taille du système augmente. Malgré tout, puisque des opérations d'insertion n'ont qu'à être effectuées que très rarement, cette perte de performance ne se révèle pas dommageable à la survie du système à long terme. La relation entre l'augmentation du temps requis pour chaque entrée dans le SGBDR et la taille de la table mesures est à peu près linéaire, avec un coût d'environ un millième de seconde par entrée à chaque passage vers un système de plus grande taille. Cette augmentation peut sembler insignifiante à petite échelle, mais pour un ensemble de données de 10 000 000 d'entrées, le temps d'exécution de la procédure d'insertion augmente de plus de trois heures. Néanmoins, l'augmentation de ce temps d'exécution reste raisonnable compte tenu de la rareté des opérations d'insertion. Ainsi, le système devrait pouvoir absorber sans problèmes les données associées aux phases subséquentes.

Opérations de sélection

Des opérations de sélection doivent être effectuées à chaque fois qu'un accès aux données est fait. Bien qu'elles soient moins coûteuses en temps de calcul que les opérations d'insertion, elles sont exécutées beaucoup plus fréquemment et il est intéressant d'analyser le comportement du système lorsque la taille des tables augmente.

Afin de procéder à une analyse complète, deux méthodes d'évaluation des performances seront utilisées, réutilisant certains des programmes permettant de produire des tableaux présentés à la Section 5.1.3. Dans un premier temps, la production d'un tableau statistique de qualité de l'air sera faite. Cette opération repose sur une seule requête, qui permet d'obtenir l'ensemble des données pour un paramètre et de les utiliser afin de produire les tableaux présentés à la Section 5.1.3. Afin de normaliser l'opération pour tous les systèmes, le tableau sera produit pour la période de référence pré-ouverture (24 mai 2010 au 23 mai 2011), pour le paramètre O_3 et ses trois critères sur une, huit et vingt-quatre heures. Les moyennes de temps requis pour trois exécutions sont présentées au Tableau 5.9. Les valeurs obtenues démontrent bien que les temps d'exécution sont constants, et ce peu importe la taille du système, en raison des indexations implantées dans le système.

Tableau 5.9: Performance du système lors de sélection simple

Taille du système	Temps d'exécution (s)
0x	N/A
1x	18.47
2x	19.09
3x	19.03
4x	18.81
5x	18.85

Il est aussi intéressant d'évaluer les performances du système lors d'opérations présentant des requêtes nombreuses. Le script produisant le nombre de dépassements à une norme pour chaque jour utilise une requête pour accéder à chaque entrée individuellement, ce qui fait en sorte qu'il représente une bonne procédure d'étalonnage pour vérifier la perte de performance lors de requêtes multiples. Dans le but de normaliser la procédure, les statistiques de performance seront produites pour les six tailles de système, pour la période de référence pré ouverture et le paramètre O_3, sur ses trois normes mentionnées précédemment. Les résultats de l'exécution sont présentés au Tableau 5.10.

Tableau 5.10: Performance du système lors de sélections complexes

Taille du système	Temps d'exécution (s)
0x	N/A
1x	595.41
2x	579.36
3x	604.15
4x	588.66
5x	587.69

Il ressort des deux tableaux présentés précédemment que le système ne subit pratiquement aucune réduction de performance dans les tests effectués, et ce, peu importe la complexité des requêtes à effectuer pour obtenir les données. Le maintien des temps d'exécution laisse croire que l'utilisation d'index rend le SGBDR apte à soutenir l'ensemble des données qui seront à disposition dans les différentes phases de ce projet, même si les cadres spatiaux et temporels étaient appelés à s'élargir.

Afin de mesurer les bénéfices associés à l'utilisation des index, le test de performance de sélection simple a été repris sur les cinq tables "mesures" utilisées précédemment après que les index de chacune d'elles aient été retirés. Les temps d'exécution requis pour ces cinq tests sont présentés au Tableau 5.11. Ce tableau illustre les pertes de performance importantes associées au retrait des index. En effet, les temps d'exécution sont multipliés par plus de 11 dans le meilleur des cas, jusqu'à 37 fois plus dans le cas de la table représentant cinq fois la taille des ensembles de données fournis. Sans l'utilisation des index, les temps d'exécution des requêtes sont donc directement proportionnels à la taille de la base de données. En plus d'augmenter le temps d'exécution des requêtes à la base de données, la désactivation des index entraîne une beaucoup plus grande utilisation des ressources systèmes pour obtenir les mêmes données. Ce dernier test de performance illustre bien les bénéfices associés à une planification minutieuse de la structure du schéma, surtout lorsque la taille de la base de données augmente.

Tableau 5.11: Performance du système lors de sélections simples (sans index)

Taille du système	Temps d'exécution (s)	Temps d'exécution (par rapport aux résultats du Tableau 5.9)
0x	N/A	N/A
1x	214.47	11.37x
2x	365.29	19.37x
3x	526.87	27.95x
4x	648.63	34.41x
5x	701.12	37.19x

CHAPITRE 6

CONCLUSION

6.1 Synthèse des travaux

Les travaux relatés dans ce mémoire visaient principalement à développer un système d'information intégré pouvant soutenir des analyses sur les impacts associés à la mise en service du pont Olivier-Charbonneau reliant l'île Jésus et l'île de Montréal. La revue de la littérature a démontré la pertinence de procéder à de telles analyses, celles-ci étant rarement accomplies, mais aussi l'intérêt de procéder au développement du système d'information.

Le système d'information développé permet non seulement de stocker l'ensemble des informations disparates fournies, mais aussi de procéder à des normalisations et de créer des contraintes permettant d'assurer la qualité et la cohérence des données stockées. Les normalisations permettent en outre d'assurer un accès simple aux données et de résoudre un grand nombre de problématiques identifiées qui rendaient nécessaire l'utilisation de documentation spécifique à chaque fichier. L'implantation du système, dans un processus itératif, a permis d'en valider la versatilité et l'extensibilité, des adaptations pouvant être faites très rapidement pour combler de nouveaux besoins identifiés.

La production de plusieurs exemples de programmes automatisés faisant l'exploitation des données stockées a permis de s'assurer de la viabilité du système afin de soutenir l'ensemble des analyses pour toute la durée du mandat de recherche sur la nouvelle infrastructure. Les technologies choisies répondent aussi aux besoins en ce qui a trait à la capacité de plusieurs intervenants de faire un usage distant des données accumulées et validées.

La contribution de ce projet se fait donc sous deux angles principaux, soit la validation de la pertinence de procéder aux étapes de conception et de traitements automatisés de données afin de soutenir les analyses pour un projet d'envergure, mais aussi les bénéfices liés à la structuration et à la normalisation d'ensembles de données variés dans le but d'automatiser des étapes à répéter tout au long d'une étude en plusieurs phases.

À tous ces égards, la solution développée permet non seulement d'assurer un soutien technique aux mandats d'analyse touchant la mise en service du pont Olivier-Charbonneau, mais établit aussi des bases méthodologiques et techniques en ce qui a trait au développement d'un système d'information spécifique à des données de transport. La qualité des données obtenues et les diverses problématiques qui y existent semblent indiquer que la gestion de données par les différents organismes en cause apporte des problèmes importants à leur utili-

sation et bénéficierait du développement d'un système d'information tel que celui développé dans le cadre de ce projet. La réalisation de ce projet a d'ailleurs été fortement ralentie par la multiplication des problématiques spécifiques associées à chaque ensemble de données, en plus des formats changeants. Il découle de cette situation qu'un très grand nombre d'heures sont perdues à réparer des problèmes de nature élémentaires et la capacité à obtenir des résultats valides dans un délai raisonnable est grandement compromise par cette situation. Par ailleurs, les formats de données extrêmement variables obtenus rendent aussi plus complexe l'intégration, ce qui vient encore une fois allonger les délais liés à la mise en place du système d'information. Si les problèmes structurels dans les jeux de données peuvent être résolus lors de l'utilisation des résultats et les multiples formats gérés au même moment, il reste que l'application de meilleurs processus en amont viendrait grandement simplifier et accélérer les processus d'intégration et d'exploitation de l'information.

6.2 Limitations de la solution proposée

Bien que la solution développée réponde bien aux objectifs énoncés, elle présente certaines limitations. Par exemple, les processus automatisés d'insertion des données gèrent bien les informations fournies, mais doivent être modifiés si jamais des changements au schéma devaient subvenir. De nouveaux ensembles de données risquent aussi d'être accompagnés par de nouvelles problématiques ponctuelles que le système n'est pas nécessairement apte à gérer dans son état actuel. Par ailleurs, toute intervention humaine risque d'amener des problèmes quant au non-respect des normes établies, ou encore en ajoutant des doublons qui viennent réduire la cohérence des données stockées. Si le système peut bien gérer des doublons formellement énoncés, il n'est pas à l'abri d'erreurs comme des variations simples dans le nom des stations ou des paramètres. Néanmoins, la méthode de développement itérative utilisée jusqu'à maintenant a permis de répondre aux changements et aux sources d'erreurs potentielles et devrait continuer à permettre d'adapter le système en fonction des besoins.

L'utilisation d'un système d'information intégré et des outils avancés choisis présentent aussi d'autres problèmes, notamment une courbe d'apprentissage importante. En effet, certains des concepts, comme la représentation hiérarchique des stations, peuvent se révéler compliqués pour des non-initiés. Dans le cadre de travaux simples basés sur des sous-ensembles de données très petits, la reconstitution des données via le regroupement de différents types d'objets et en tenant compte de principes hiérarchiques peut s'avérer inutilement complexe.

L'administration doit aussi être accomplie par quelqu'un possédant une connaissance approfondie du système. Par conséquent, le système court le risque d'être laissé à l'abandon en l'absence d'une ou des personnes qualifiées pour en faire la maintenance, garder les informa-

tions stockées à jour ou encore en étendre la portée spatiale ou temporelle.

Finalement, bien que le système fonctionne bien dans le paradigme de stations, paramètres et mesures, l'ajout de données extérieures à ce paradigme, comme des données zonales, est un processus qui demande une série d'interventions complexes ou encore la réorganisation d'une partie des données déjà contenues. De tels besoins peuvent faire en sorte de rendre le système en entier obsolète si la tâche de restructuration est jugée trop importante par rapport aux bénéfices.

6.3 Perspectives

Le système étant conçu pour accumuler l'ensemble des données fournies dans le cadre des mandats d'évaluation actuels, et ayant la capacité de croître sans présenter de réduction de performances, il devrait pouvoir sans problème servir de base aux analyses de l'équipe de recherche, notamment pour la production de tableaux et de graphiques, et ce, pour toute la période de suivi.

Ces analyses produites jusqu'à maintenant à partir des données stockées sont assez peu nombreuses, se limitant à la production de tableaux et de graphiques. Toutefois, le système présente des capacités pratiquement illimitées grâce à la structuration et à la normalisation des données. Il devrait notamment y avoir des possibilités importantes d'opération de data-mining sur les données stockées afin de pouvoir identifier des tendances dans le cadre d'analyses multivariées.

Les plus grandes potentialités du système se présentent toutefois sous la possibilité d'en faire l'utilisation pour soutenir d'autres projets de recherche. En effet, la masse de données stockées pourrait se révéler intéressante pour une multitude de projets, qui pourraient alors faire un usage direct du système d'information, ou encore s'y arrimer en ajoutant d'autres sources de données pertinentes en profitant de ses capacités d'extensibilité. Cette utilisation viendrait potentiellement offrir des gains de temps du point de vue de l'organisation de l'information et ainsi accélérer l'obtention de résultats concrets.

RÉFÉRENCES

AMBLER, S. (2012). *Agile Database Techniques: Effective Strategies for the Agile Software Developer*. Wiley.

AMBLER, S. W. et SADALAGE, P. J. (2006). *Refactoring Databases: Evolutionary Database Design*. Addison-Wesley Signature Series. Pearson Education.

BAIN, R. (2009). Error and optimism bias in toll road traffic forecasts. *Transportation*, 36, 469–482.

BANISTER, D. (2005). *Unsustainable Transport: City Transport in the New Century*. Transport, Development and Sustainability Series. Taylor & Francis.

BARRON, K., COLLINS, J., DERR, R. et JACOBSON, L. (2004). The future of travel time data—a paradigm shift.

BONSALL, P. et O'FLAHERTY, C. (1997). Observational traffic surveys. *Transportation Planning and Traffic Engineering*.

BRACKETT, M. H. (1996). *The data warehouse challenge: taming data chaos*. Wiley computer publishing. Wiley.

BRAESS, D., NAGURNEY, A. et WAKOLBINGER, T. (2005). On a Paradox of Traffic Planning. *Transportation Science*, 39, 446–450.

BROWNSTONE, D. et SMALL, K. (2005). Valuing time and reliability: assessing the evidence from road pricing demonstrations. *Transportation Research Part A: Policy and Practice*, 39, 279–293.

CARRION, C. et LEVINSON, D. (2012). A Model of Bridge Choice Across the Mississippi River in Minneapolis. *Network Reliability in Practice*.

CERVERO, R. (2003). Road expansion, urban growth, and induced travel: A path analysis. *Journal of the American Planning Association*.

CHEN, C. et PETTY, K. (2001). Freeway performance measurement system: mining loop detector data. *Transportation Research Record: Journal of the Transportation Research Board*, 1748, 96–102.

CHEN, M., HAN, J. et YU, P. (1996). Data mining: an overview from a database perspective. *Knowledge and Data Engineering, IEEE Transactions on*, 8, 866–883.

CLEGG, R. (2007). Empirical Studies on Road Traffic Response to Capacity Reduction. *Transportation and Traffic Theory*.

DAHLGREN, J. (1998). *Methodologies for Assessing the Impacts of Highway Capacity Enhancements on Travel Behavior*.

DAHLGREN, J. (2001). How the Reconstruction of I-880 Affected Travel Behavior.

DAHLGREN, J., GARCIA, R. et TURNER, S. (2001). Completing the Circle: Using Archived Operations Data to Better Link Decisions to Performance.

DAHLGREN, J. et STATION, R. F. (2003). The benefits of accommodating latent demand. *Transportation Research Board (TRB) 2003 Annual Meeting CD-ROM*.

DANCZYK, A., LIU, H. et LEVINSON, D. (2010). Unexpected cause, unexpected effect: Empirical observations of Twin Cities traffic behavior after the I-35W Bridge collapse and reopening. *Transportation*.

DECORLA-SOUZA, P. et COHEN, H. (1999). Accounting for induced travel in evaluation of urban highway expansion. *Sixth National Conference on Transportation Planning for Small and Medium-Sized Communities*.

DUEKER, K. et BUTLER, J. (2000). A geographic information system framework for transportation data sharing. *Transportation Research Part C: Emerging Technologies*, 8, 13–36.

ELMASRI, R. A. et NAVATHE, S. B. (2011). *Fundamentals of Database Systems*. ADDISON WESLEY Publishing Company Incorporated.

ETCHES, A., CLARAMUNT, C., BARGIELA, A. et KOSONEN, I. (1998). An integrated temporal GIS model for traffic systems. *Innovations in GIS*.

FLYVBJERG, B. (2005). Measuring inaccuracy in travel demand forecasting: methodological considerations regarding ramp up and sampling. *Transportation Research Part A: Policy and Practice*, 39, 522–530.

FLYVBJERG, B., BRUZELIUS, N. et ROTHENGATTER, W. (2003). *Megaprojects and Risk: An Anatomy of Ambition*. Cambridge University Press.

FLYVBJERG, B., HOLM, M. et BUHL, S. (2006). Inaccuracy in traffic forecasts. *Transport Reviews*, 26, 1–24.

FRIHIDA, A., MARCEAU, D. et THERIAULT, M. (2008). Spatio-temporal object-oriented data model for disaggregate travel behavior. *Transactions in GIS*, 6.

FU, Y., LI, Z., SONG, K., QIU, Z. et MA, X. (2006). Integrated traffic management platform design based on GIS-T. *ITS Telecommunications Proceedings, 2006 6th International Conference on*, 29–32.

GIULIANO, G. et GOLOB, J. (1998). Impacts of the Northridge earthquake on transit and highway use. *Journal of transportation and statistics*, 1–20.

GOODCHILD, M. (2000). GIS and transportation: status and challenges. *GeoInformatica*.

GOODWIN, P. (1996). Empirical evidence on induced traffic. *Transportation*, 35–54.

GOODWIN, P., ATKINS, S. et CAIRNS, S. (2002). Disappearing traffic ? The story so far. *Proceedings of the ICE - Municipal Engineer*, 151, 13–22.

GOODWIN, P. et NOLAND, R. (2003). Building new roads really does create extra traffic: a response to Prakash et al. *Applied Economics*, 44, 1–19.

GUTIERREZ, J. et GÓMEZ, G. (1999). The impact of orbital motorways on intra-metropolitan accessibility: the case of Madrid's M-40. *Journal of Transport Geography*.

HANSEN, M., GILLEN, D. et DOBBINS, A. (1993). The air quality impacts of urban highway capacity expansion: Traffic generation and land use change.

HE, X., JABARI, S. et LIU, H. X. (2008). Modeling Day-to-day Trip Choice Evolution under Network Disruption.

HIGHWAYS AGENCY (2006). A34 NEWBURY BYPASS 'Five Years After' Evaluation (1998–2003). Rapport technique July, Highways Agency, London.

HUANG, Z. (2003). *Data integration for urban transport planning*.

HULTEN, G., SPENCER, L. et DOMINGOS, P. (2001). Mining time-changing data streams. *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '01*, 97–106.

JACK FAUCET ASSOCIATES (1997). *NCHRP Web Doc 3 Multimodal Transportation Planning Data:Final Report*. The National Academies Press, Washington DC.

KWON, J., COIFMAN, B. et BICKEL, P. (2000). Day-to-day travel-time trends and travel-time prediction from loop-detector data. *Transportation Research Record: Journal of the Transportation Research Board*, 1717, 1–18.

LAIRD, J. J., NELLTHORP, J. et MACKIE, P. J. (2005). Network effects and total economic impact in transport appraisal. *Transport Policy*, 12, 537–544.

LINNEKER, B. et SPENCE, N. (1996). Road transport infrastructure and regional economic development. *Journal of Transport Geography*, 4, 77–92.

MACKIE, P. et PRESTON, J. (1998). Twenty-one sources of error and bias in transport project appraisal. *Transport Policy*, 5.

MARTIN, R. C. (2003). *Agile Software Development: Principles, Patterns, and Practices*. Prentice Hall PTR, Upper Saddle River, NJ, USA.

METZ, D. (2008). The Myth of Travel Time Saving. *Transport Reviews*, 28, 321–336.

MILLER, H. (1999). Potential contributions of spatial analysis to geographic information systems for transportation (GIS-T). *Geographical Analysis*, 31.

NOLAND, R. B. (2001). Relationships between highway capacity and induced vehicle travel. *Transportation Research Part A: Policy and Practice*, 35, 47–72.

OLSEN JR, D. R. (1999). Interacting in chaos. *Interactions*, 6, 42–54.

PARTHASARATHI, P. et LEVINSON, D. (2010). Post-construction evaluation of traffic forecast accuracy. *Transport Policy*.

PFLEIDERER, R. et DIETERICH, M. (1995). New roads generate new traffic. *World Transport Policy and Practice*, 1, 29–31.

PIARC TECHNICAL COMMITTEE (2012). Worldwide situation of road pricing and assessment of its impacts. Rapport technique.

PISARSKI, A. E. (1997). Information Needs To Support State and Transportation Decision Making into the 21st Century Local. Transportation Research Board, Irvine, Californie, 1–92.

PRAKASH, A., OLIVER, I. et BALCOMBE, K. (2001). Does building new roads really create extra traffic ? Some new evidence. *Applied Economics*, 37–41.

PYLE, D. (1999). *Data Preparation for Data Mining: Text*. The Morgan Kaufmann Series in Data Management Systems Series. Morgan Kaufmann Pub.

QUIROGA, C. (2000). Performance measures and data requirements for congestion management systems. *Transportation Research Part C: Emerging Technologies*, 8, 287–306.

SCHOFER, J. L. (2007). Information Assets to Support Transportation Decision Making. Rapport technique August, Northwestern University, Kansas City.

SHAW, S. et WANG, D. (2000). Handling disaggregate spatiotemporal travel data in GIS. *GeoInformatica*.

SHORT, J. et KOPP, A. (2005). Transport infrastructure: investment and planning. Policy and research aspects. *Transport Policy*, 12, 360–367.

THILL, J. (2000). Geographic information systems for transportation in perspective. *Transportation Research Part C: Emerging Technologies*, 8, 3–12.

TRÉPANIER, M. et CHAPLEAU, R. (2001). Analyse orientée-objet et totalement désagrégée des données d'enquêtes ménages origine-destination. *Canadian Journal of Civil Engineering*, 28, 48–58.

VALSECCHI, P., CLARAMUNT, C. et PEYTCHEV, E. (1999). OSIRIS: an inter-operable system for the integration of real-time traffic data within GIS. *Computers, Environment and Urban Systems*, 23, 245–257.

WATERS, N. (1999). Transportation GIs: GIs-T. *Geographical information systems*.

XIE, F. et LEVINSON, D. (2011). Evaluating the effects of the I-35W bridge collapse on road-users in the twin cities metropolitan region. *Transportation Planning and Technology*.

ZHU, S. et LEVINSON, D. (2008). A review of research on planned and unplanned disruptions to transportation networks.

ZHU, S. et LEVINSON, D. (2010). Travels Impacts of Bridge Closures 1: Lafayette Bridge Final Report.

ZHU, S., LEVINSON, D., LIU, H. et HARDER, K. (2010). The traffic and behavioral effects of the I-35W Mississippi River bridge collapse. *Transportation research part A: policy and practice*, 44, 771–784.

ZHU, S., LEVINSON, D. M. et LIU, H. (2009). Measuring Winners and Losers from the New I-35w Mississippi River Bridge. *SSRN Electronic Journal*, 1–19.

ZHU, S., TILAHUN, N., HE, X. et LEVINSON, D. (2012). Travel Impacts and Adjustment Strategies of the Collapse and the Reopening of the I-35W Bridge. *Network Reliability in Practice*, 1–23.

ZILIASKOPOULOS, A. et WALLER, S. (2000). An Internet-based geographic information system that integrates data, models and users for transportation applications. *Transportation Research Part C: Emerging Technologies*, 8, 427–444.

ANNEXE A

Fonction d'horodatage

Cette fonction permet de créer un tableau comprenant l'ensemble des heures entre deux horodatages définis (variables temps_minimum et temps_maximum, attendues comme des valeurs texte obtenues d'un fichier csv) ainsi que des horodatages ajustés en fonction des changements d'heure (variable heure_avancee = 1 si le fichier d'origine ne tient pas compte des changements d'heure et que ceux-ci doivent être ajustés) et du besoin d'ajuster les horodatages aux normes du système d'information (variable temps_compensation).

Code A.1: Fonction de gestion des horodatages

```
1 def creer_tableau_controle(temps_minimum, temps_maximum, heure_avancee,
      temps_compensation = 0.hours)
2     format_temps = "%Y-%m-%d %H:%M:%S"
3     if temps_minimum.is_a? DateTime
4         temps_minimum = temps_minimum.strftime(format_temps)
5     end
6     if temps_maximum.is_a? DateTime
7         temps_maximum = temps_maximum.strftime(format_temps)
8     end
9     if heure_avancee == 0
10        temps_depart = DateTime.parse(temps_minimum)
11        temps_fin = DateTime.parse(temps_maximum)
12        tableau_controle = []
13        temps_actuel = temps_depart
14        i = 0
15        statut_heure_avancee_precedent = parselocaltime((temps_depart  -
              1.hour).strftime(format_temps)).dst?
16        while temps_actuel <= temps_fin do
17            temps_actuel_texte = (temps_actuel - 1.second).strftime(format_temps)
18            chaine_recherche = temps_actuel.strftime(format_temps)
19            temps_local = parselocaltime(temps_actuel_texte)
20            statut_heure_avancee = temps_local.dst?
21            tableau_controle[i] = {horodatage:temps_actuel  -  1.second +
                  temps_compensation, chaine_recherche:chaine_recherche,
                  heure_avancee:statut_heure_avancee}
22            statut_heure_avancee_precedent = statut_heure_avancee
```

Code A.1 (Suite)

```
23              temps_actuel += 1.hour
24              i += 1
25          end
26          return tableau_controle
27      else
28          firsttimedst = parselocaltime(temps_minimum).dst?
29          lasttimedst = parselocaltime(temps_maximum).dst?
30          temps_depart = DateTime.parse(temps_minimum)
31          if firsttimedst
32              temps_depart += 1.hour
33          end
34          temps_fin = DateTime.parse(temps_maximum)
35          if lasttimedst
36              temps_fin += 1.hour
37          end
38          tableau_controle = []
39          temps_actuel = temps_depart
40          i = 0
41          statut_heure_avancee_precedent = parselocaltime((temps_depart  -  1.hour +
                temps_compensation).strftime(format_temps)).dst?
42          while temps_actuel <= temps_fin do
43              temps_actuel_texte = (temps_actuel - 1.second).strftime(format_temps)
44              temps_local = parselocaltime(temps_actuel_texte)
45              statut_heure_avancee = (temps_local + temps_compensation).dst?
46              if statut_heure_avancee and statut_heure_avancee_precedent
47                  chaine_recherche = (temps_actuel  -  1.hour).strftime(format_temps)
48              elsif statut_heure_avancee and !statut_heure_avancee_precedent
49                  chaine_recherche = nil
50              else
51                  chaine_recherche = temps_actuel.strftime(format_temps)
52              end
53              tableau_controle[i] = {horodatage:temps_actuel  -  1.second +
                    temps_compensation, chaine_recherche:chaine_recherche,
                    heure_avancee:statut_heure_avancee}
54              statut_heure_avancee_precedent = statut_heure_avancee
55              temps_actuel += 1.hour
56              i += 1
57          end
58          return tableau_controle
59      end
60  end
```

ANNEXE B

Fonction d'importation des données météo

Code B.1: Fonction d'importation des donnée météo

```
def import_fichier_ascii(chemin)
    fichier = File.open(chemin)
    lignes = []
    fichier.each_line {|f| lignes << f}
    sortie = []
    lignes.each {|ligne|
        station = ligne[0..6]
        date = ligne[8..11].to_s + "-" +ligne[13..14].to_s + "-" +
            ligne[16..17].to_s
        parametre = ligne[19..23].to_i
        outlines = (0..23).step
        outlines.each {|outline|
            id = outline.to_i
            idr = ((id * 9) + 27)..((id * 9) + 32)
            heure = sprintf('%02.f', id) + ":" + sprintf('%02.f', 0) + ":" +
                sprintf('%02.f', 0)
            horodatage = date + "␣" + heure
            horodatage = DateTime.parse(horodatage)
            qual = (id * 9) + 34
            qualite = ligne[qual].strip
            mesure = ligne[idr].strip
            qualite = "" if qualite.nil?
            mesure = "" if mesure.nil?
            if qualite.empty?
                qualite = nil
            else
                qualite = qualite.to_i
            end
            if mesure.empty?
                mesure = nil
            else
                mesure = mesure.to_f
            end
```

Code B.1 (Suite)

```
32              sortie << {station:station, parametre:parametre, horodatage:horodatage,
                    mesure:mesure, qualite:qualite}
33          }
34      }
35      return sortie
36  end
```

ANNEXE C

Stations et mesures associées

Tableau C.1: Nombre de données et types associées à chaque station créée

Station	Type	Compte
L - H Lafontaine (S2)	Qualité de l'air - Molécules	43 447
Châteauneuf (S1)	Qualité de l'air - Molécules	68 744
Autoroute 25 (C1)	Qualité de l'air - Molécules / COV	94 859
Perras (C2)	Qualité de l'air - Molécules / COV	91 412
Roger Lortie (N1)	Qualité de l'air - Molécules	60 876
Lévesque (N2)	Qualité de l'air - Molécules	70 482
Chénier,Anjou (7)	Qualité de l'air - Molécules / COV	79 516
Parc Pilon - Montréal - Nord (29)	Qualité de l'air - Molécules	125 931
Rivière - des - prairies (55)	Qualité de l'air - Molécules / COV	82 587
Montréal - Autoroute - 25	Météo	65 903
Pont Viau voie 1	Circulation - Classifiées	14 454
Pont Viau voie 2	Circulation - Classifiées	14 454
Pont Viau voie 3	Circulation - Classifiées	10 233
Pont Viau voie 6	Circulation - Classifiées	14 445
Pont Viau voie 7	Circulation - Classifiées	14 445
Tunnel Louis - Hippolyte - La Fontaine voie 1	Circulation - Classifiées	50 725
Tunnel Louis - Hippolyte - La Fontaine voie 2	Circulation - Classifiées	82 501
Suite à la page suivante		

Tableau C.1: Nombre de données et types associées à chaque station créée (suite)

Station	Type	Compte
Tunnel Louis - Hippolyte - La Fontaine voie 3	Circulation - Classifiées	65 097
Tunnel Louis - Hippolyte - La Fontaine voie 6	Circulation - Classifiées	71 799
Tunnel Louis - Hippolyte - La Fontaine voie 7	Circulation - Classifiées	85 125
Tunnel Louis - Hippolyte - La Fontaine voie 8	Circulation - Classifiées	72 614
Pont Lepage voie 1	Circulation - Classifiées	215 989
Pont Lepage voie 2	Circulation - Classifiées	464 696
Pont Lepage voie 3	Circulation - Classifiées	320 087
Pont Lepage voie 6	Circulation - Classifiées	259 178
Pont Lepage voie 7	Circulation - Classifiées	497 199
Pont Lepage voie 8	Circulation - Classifiées	295 272
Pont Papineau voie 1	Circulation - Classifiées	736 789
Pont Papineau voie 2	Circulation - Classifiées	857 665
Pont Papineau voie 3	Circulation - Classifiées	617 877
Pont Papineau voie 6	Circulation - Classifiées	537 720
Pont Papineau voie 7	Circulation - Classifiées	681 650
Pont Papineau voie 8	Circulation - Classifiées	825 578
Pont Papineau voie 9	Circulation - Classifiées	608 164
Pont Pie IX voie 1	Circulation - Classifiées	292 501
Pont Pie IX voie 2	Circulation - Classifiées	445 306
Pont Pie IX voie 3	Circulation - Classifiées	296 757
Pont Pie IX voie 6	Circulation - Classifiées	366 894
Pont Pie IX voie 7	Circulation - Classifiées	424 777
Pont Pie IX voie 8	Circulation - Classifiées	301 186
Pont Charles - De Gaulle voie 1	Circulation - Classifiées	17 373
Pont Charles - De Gaulle voie 2	Circulation - Classifiées	17 361
Pont Charles - De Gaulle voie 3	Circulation - Classifiées	16 196
Pont Charles - De Gaulle voie 6	Circulation - Classifiées	20 840
Pont Charles - De Gaulle voie 7	Circulation - Classifiées	20 840
Pont Charles - De Gaulle voie 8	Circulation - Classifiées	19 101
Pont Le Gardeur voie 1	Circulation - Classifiées	370 865

Suite à la page suivante

Tableau C.1: Nombre de données et types associées à chaque station créée (suite)

Station	Type	Compte
Pont Le Gardeur voie 2	Circulation - Classifiées	386 209
Pont Le Gardeur voie 3	Circulation - Classifiées	91 369
Pont Le Gardeur voie 6	Circulation - Classifiées	430 845
Pont Le Gardeur voie 7	Circulation - Classifiées	359 563
A - 40 près de Galeries d'Anjou	Circulation - Comptages	1 837
A - 40 près de Galeries d'Anjou	Circulation - Comptages	1 894
A - 25 au Nord de Montée Saint - François	Circulation - Comptages	374
A - 25 au Nord de Montée Saint - François	Circulation - Comptages	564
A - 640 Entre Montée Dumais et Montée des Pionniers	Circulation - Comptages	17 151
A - 640 Entre Montée Dumais et Montée des Pionniers	Circulation - Comptages	16 729
Montée Masson au Nord de Avenue Marcel Villeneuve	Circulation - Comptages / Vitesse moyenne	1 334
Montée Masson au Nord de Avenue Marcel Villeneuve	Circulation - Comptages / Vitesse moyenne	1 334
Bretelle Boulevard des Milles - Îles vers A - 25S	Circulation - Comptages	909
Bretelle A - 25S vers Boulevard des Milles - Îles	Circulation - Comptages	909
Bretelle Boulevard des Milles - Îles vers A - 25N	Circulation - Comptages	909
Bretelle A - 25N vers Boulevard des Milles - Îles	Circulation - Comptages	911
Bretelle A - 25S Sortie Sherbrooke	Circulation - Comptages	2 209
Bretelle A - 25S Sortie Souligny	Circulation - Comptages	1 821
Intersection Boulevard Henri - Bourassa Est et Boulevard Louis - H Lafontaine approche Nord direction Ouest	Circulation - Comptages	292
Intersection Boulevard Henri - Bourassa Est et Boulevard Louis - H Lafontaine approche Nord direction Sud	Circulation - Comptages	292
Intersection Boulevard Henri - Bourassa Est et Boulevard Louis - H Lafontaine approche Nord direction Est	Circulation - Comptages	292

Suite à la page suivante

Tableau C.1: Nombre de données et types associées à chaque station créée (suite)

Station	Type	Compte
Intersection Boulevard Henri - Bourassa Est et Boulevard Louis - H Lafontaine approche Est direction Nord	Circulation - Comptages	292
Bretelle A - 25S Sortie Beaubien	Circulation - Comptages	657
A - 40 près de Galeries d'Anjou voie 2	Circulation - Comptages	1 926
A - 40 près de Galeries d'Anjou voie 3	Circulation - Comptages	1 926
A - 40 près de Galeries d'Anjou voie 6	Circulation - Comptages	1 537
A - 40 près de Galeries d'Anjou voie 7	Circulation - Comptages	1 537
A - 40 près de Galeries d'Anjou voie 8	Circulation - Comptages	1 537
A - 40 Près de Langelier voie 1	Circulation - Comptages	387
A - 40 Près de Langelier voie 2	Circulation - Comptages	387
A - 40 Près de Langelier voie 3	Circulation - Comptages	387
A - 40 Près de Langelier voie 6	Circulation - Comptages	387
A - 40 Près de Langelier voie 7	Circulation - Comptages	387
A - 40 Près de Langelier voie 8	Circulation - Comptages	387
Intersection Boulevard Henri - Bourassa Est et Boulevard Louis - H Lafontaine approche Est direction Ouest	Circulation - Comptages	292
Intersection Boulevard Henri - Bourassa Est et Boulevard Louis - H Lafontaine approche Est direction Sud	Circulation - Comptages	292
Intersection Boulevard Henri - Bourassa Est et Boulevard Louis - H Lafontaine approche Sud direction Est	Circulation - Comptages	292
Intersection Boulevard Henri - Bourassa Est et Boulevard Louis - H Lafontaine approche Sud direction Nord	Circulation - Comptages	292
Intersection Boulevard Henri - Bourassa Est et Boulevard Louis - H Lafontaine approche Sud direction Ouest	Circulation - Comptages	292
Suite à la page suivante		

Tableau C.1: Nombre de données et types associées à chaque station créée (suite)

Station	Type	Compte
Intersection Boulevard Henri - Bourassa Est et Boulevard Louis - H Lafontaine approche Ouest direction Sud	Circulation - Comptages	292
Intersection Boulevard Henri - Bourassa Est et Boulevard Louis - H Lafontaine approche Ouest direction Est	Circulation - Comptages	292
Intersection Boulevard Henri - Bourassa Est et Boulevard Louis - H Lafontaine approche Ouest direction Nord	Circulation - Comptages	292
Bretelle Montée Masson vers A - 25N	Circulation - Comptages	144
Avenue Marcel Villeneuve (1km à l'est de Avenue Roger - Lortie)	Circulation - Comptages / Vitesse moyenne	1 326
Avenue Marcel Villeneuve (1km à l'est de Avenue Roger - Lortie)	Circulation - Comptages / Vitesse moyenne	1 326
Boulevard Lévesque Est (300m à l'est de Avenue Roger - Lortie)	Circulation - Comptages / Vitesse moyenne	504
Boulevard Lévesque Est (300m à l'est de Avenue Roger - Lortie)	Circulation - Comptages / Vitesse moyenne	505
Boulevard Lévesque Est (400m à l'ouest de Avenue Roger - Lortie)	Circulation - Comptages / Vitesse moyenne	526
Boulevard Lévesque Est (400m à l'ouest de Avenue Roger - Lortie)	Circulation - Comptages / Vitesse moyenne	524
Boulevard Louis - H - Lafontaine au Nord de boulevard Maurice Duplessis	Circulation - Comptages	480
Boulevard Louis - H - Lafontaine au Nord de boulevard Maurice Duplessis	Circulation - Comptages	480
A - 25 au Sud de A - 40	Circulation - Comptages	1 864
A - 25 au Sud de A - 40	Circulation - Comptages	1 872
A - 25 au Sud de de Bombardier	Circulation - Comptages	2 000
A - 25 au Sud de de Bombardier	Circulation - Comptages	2 000
Avenue Roger - Lortie (Intersection voie CP) voie 1	Circulation - Comptages	955
Suite à la page suivante		

Tableau C.1: Nombre de données et types associées à chaque station créée (suite)

Station	Type	Compte
A - 25 au Sud de A - 40 voie 1	Circulation - Comptages / Vitesse moyenne	15 600
A - 25 au Sud de A - 40 voie 2	Circulation - Comptages / Vitesse moyenne	15 600
A - 25 au Sud de A - 40 voie 3	Circulation - Comptages / Vitesse moyenne	15 380
A - 25 au Sud de A - 40 voie 1	Circulation - Comptages / Vitesse moyenne	119 560
A - 25 au Sud de A - 40 voie 2	Circulation - Comptages / Vitesse moyenne	119 560
A - 25 au Sud de A - 40 voie 3	Circulation - Comptages / Vitesse moyenne	119 560
A - 25 au Sud de A - 40 voie 4	Circulation - Comptages / Vitesse moyenne	119 560
A - 25 au Sud de de Bombardier voie 1	Circulation - Comptages	1 325
A - 25 au Sud de de Bombardier voie 2	Circulation - Comptages	1 325
A - 25 au Sud de de Bombardier voie 4	Circulation - Comptages	1 325
A - 25 au Sud de de Bombardier voie 6	Circulation - Comptages	1 325
A - 25 au Sud de de Bombardier voie 7	Circulation - Comptages	1 325
A - 25 au Sud de de Bombardier voie 9	Circulation - Comptages	1 325
Intersection Boulevard Henri - Bourassa Est et Boulevard Lacordaire approche Nord direction Ouest	Circulation - Comptages	292
Intersection Boulevard Henri - Bourassa Est et Boulevard Lacordaire approche Nord direction Sud	Circulation - Comptages	292
Intersection Boulevard Henri - Bourassa Est et Boulevard Lacordaire approche Nord direction Est	Circulation - Comptages	292
Intersection Boulevard Henri - Bourassa Est et Boulevard Lacordaire approche Est direction Nord	Circulation - Comptages	292
Intersection Boulevard Henri - Bourassa Est et Boulevard Lacordaire approche Est direction Ouest	Circulation - Comptages	292
Suite à la page suivante		

Tableau C.1: Nombre de données et types associées à chaque station créée (suite)

Station	Type	Compte
Intersection Boulevard Henri - Bourassa Est et Boulevard Lacordaire approche Est direction Sud	Circulation - Comptages	292
Intersection Boulevard Henri - Bourassa Est et Boulevard Lacordaire approche Sud direction Est	Circulation - Comptages	292
Intersection Boulevard Henri - Bourassa Est et Boulevard Lacordaire approche Sud direction Nord	Circulation - Comptages	292
Intersection Boulevard Henri - Bourassa Est et Boulevard Lacordaire approche Sud direction Ouest	Circulation - Comptages	292
Intersection Boulevard Henri - Bourassa Est et Boulevard Lacordaire approche Ouest direction Sud	Circulation - Comptages	292
Intersection Boulevard Henri - Bourassa Est et Boulevard Lacordaire approche Ouest direction Est	Circulation - Comptages	292
Intersection Boulevard Henri - Bourassa Est et Boulevard Lacordaire approche Ouest direction Nord	Circulation - Comptages	292
Intersection Boulevard Henri - Bourassa Est et Boulevard Saint - Michel approche Nord direction Ouest	Circulation - Comptages	32
Intersection Boulevard Henri - Bourassa Est et Boulevard Saint - Michel approche Nord direction Sud	Circulation - Comptages	32
Intersection Boulevard Henri - Bourassa Est et Boulevard Saint - Michel approche Nord direction Est	Circulation - Comptages	32
Intersection Boulevard Henri - Bourassa Est et Boulevard Saint - Michel approche Est direction Nord	Circulation - Comptages	32
Intersection Boulevard Henri - Bourassa Est et Boulevard Saint - Michel approche Est direction Ouest	Circulation - Comptages	32
Intersection Boulevard Henri - Bourassa Est et Boulevard Saint - Michel approche Est direction Sud	Circulation - Comptages	32

Suite à la page suivante

Tableau C.1: Nombre de données et types associées à chaque station créée (suite)

Station	Type	Compte
Intersection Boulevard Henri - Bourassa Est et Boulevard Saint - Michel approche Sud direction Est	Circulation - Comptages	32
Intersection Boulevard Henri - Bourassa Est et Boulevard Saint - Michel approche Sud direction Nord	Circulation - Comptages	32
Intersection Boulevard Henri - Bourassa Est et Boulevard Saint - Michel approche Sud direction Ouest	Circulation - Comptages	32
Intersection Boulevard Henri - Bourassa Est et Boulevard Saint - Michel approche Ouest direction Sud	Circulation - Comptages	32
Intersection Boulevard Henri - Bourassa Est et Boulevard Saint - Michel approche Ouest direction Est	Circulation - Comptages	32
Intersection Boulevard Henri - Bourassa Est et Boulevard Saint - Michel approche Ouest direction Nord	Circulation - Comptages	32
Intersection Boulevard Henri - Bourassa Est et Boulevard PieIX approche Sud direction Est	Circulation - Comptages	286
Intersection Boulevard Henri - Bourassa Est et Boulevard PieIX approche Sud direction Ouest	Circulation - Comptages	28
Intersection Boulevard Henri - Bourassa Est et Boulevard PieIX approche Ouest direction Nord	Circulation - Comptages	28
Intersection Boulevard Henri - Bourassa Est et Boulevard PieIX approche Nord direction Ouest	Circulation - Comptages	260
Intersection Boulevard Henri - Bourassa Est et Boulevard PieIX approche Ouest direction Sud	Circulation - Comptages	260
Intersection Boulevard Henri - Bourassa Est et Boulevard PieIX approche Est direction Nord	Circulation - Comptages	260
Intersection Boulevard Henri - Bourassa Est et A - 19 approche Nord direction Ouest	Circulation - Comptages	260

Suite à la page suivante

Tableau C.1: Nombre de données et types associées à chaque station créée (suite)

Station	Type	Compte
Intersection Boulevard Henri - Bourassa Est et A - 19 approche Nord direction Sud	Circulation - Comptages	260
Intersection Boulevard Henri - Bourassa Est et A - 19 approche Nord direction Est	Circulation - Comptages	260
Intersection Boulevard Henri - Bourassa Est et A - 19 approche Est direction Nord	Circulation - Comptages	260
Intersection Boulevard Henri - Bourassa Est et A - 19 approche Est direction Ouest	Circulation - Comptages	260
Intersection Boulevard Henri - Bourassa Est et A - 19 approche Est direction Sud	Circulation - Comptages	260
Intersection Boulevard Henri - Bourassa Est et A - 19 approche Sud direction Est	Circulation - Comptages	260
Intersection Boulevard Henri - Bourassa Est et A - 19 approche Sud direction Nord	Circulation - Comptages	260
Intersection Boulevard Henri - Bourassa Est et A - 19 approche Sud direction Ouest	Circulation - Comptages	260
Intersection Boulevard Henri - Bourassa Est et A - 19 approche Ouest direction Sud	Circulation - Comptages	260
Intersection Boulevard Henri - Bourassa Est et A - 19 approche Ouest direction Est	Circulation - Comptages	260
Intersection Boulevard Henri - Bourassa Est et A - 19 approche Ouest direction Nord	Circulation - Comptages	260
A - 25N Échangeur Anjou	Circulation - Comptages	660
Bretelle A - 40E vers A - 25N	Circulation - Comptages	2 873
Bretelle A - 40E vers A - 25S	Circulation - Comptages	112
A - 40E Échangeur Anjou	Circulation - Comptages	660
A - 40E - Approche bretelle A - 40E vers A - 25N	Circulation - Comptages	660
Bretelle A - 40O vers A - 25N	Circulation - Comptages	280
Bretelle A - 40O vers A - 25S	Circulation - Comptages	112
Bretelle A - 25N vers A - 40E	Circulation - Comptages	112
Bretelle A - 25N vers A - 40O	Circulation - Comptages	112

Suite à la page suivante

Tableau C.1: Nombre de données et types associées à chaque station créée (suite)

Station	Type	Compte
Bretelle A - 25S vers A - 40O	Circulation - Comptages	1 928
Bretelle A - 25S vers A - 40E	Circulation - Comptages	3 248
A - 40O Échangeur Anjou	Circulation - Comptages	660
A - 40E - Approche bretelle A - 40E vers A - 25N	Circulation - Comptages	240
A - 40E - Approche bretelle A - 40E vers A - 25N	Circulation - Comptages	240
A - 40O Échangeur Anjou	Circulation - Comptages	240
A - 40O Échangeur Anjou	Circulation - Comptages	240
Bretelle A - 25N vers A - 40O	Circulation - Comptages	240
Bretelle A - 25S vers A - 40O	Circulation - Comptages	240
Autoroute 25 au nord du Boulevard Yves - Prévost voie 1	Circulation - Comptages	336
Autoroute 25 au sud du Boulevard Wilfrid - Pelletier voie 2	Circulation - Comptages	336
Autoroute 25 au nord du Boulevard Yves - Prévost voie 2	Circulation - Comptages	336
Autoroute 25 au nord du Boulevard Yves - Prévost voie 3	Circulation - Comptages	336
Autoroute 25 au sud du Boulevard Wilfrid - Pelletier voie 3	Circulation - Comptages	336
Bretelle d'entrée Autoroute 25 - N au sud du Boulevard Wilfrid - Pelletier	Circulation - Comptages	336
Bretelle de sortie Autoroute 25 - N près du Boulevard Yves - Prévost	Circulation - Comptages	336
A - 40 près de Galeries d'Anjou voie 1	Circulation - Comptages	1 926
Avenue Roger - Lortie (Intersection voie CP) voie 9	Circulation - Comptages	955
Autoroute 25 au sud du Boulevard Wilfrid - Pelletier voie 1	Circulation - Comptages	336
Autoroute 25 au sud du Boulevard Wilfrid - Pelletier voie 6	Circulation - Comptages	336
Autoroute 25 au nord du Boulevard Yves - Prévost voie 6	Circulation - Comptages	336

Suite à la page suivante

Tableau C.1: Nombre de données et types associées à chaque station créée (suite)

Station	Type	Compte
Autoroute 25 au nord du Boulevard Yves - Prévost voie 7	Circulation - Comptages	336
Autoroute 25 au sud du Boulevard Wilfrid - Pelletier voie 7	Circulation - Comptages	336
Autoroute 25 au nord du Boulevard Yves - Prévost voie 8	Circulation - Comptages	336
Autoroute 25 au sud du Boulevard Wilfrid - Pelletier voie 8	Circulation - Comptages	336

ANNEXE D

Graphique de circulation

Code D.1: Script de production de graphique de circulation

```
g = Gruff::Line.new
g.title = "Routes"
nom_stations = ["ROUTE", "ROUTE", "ROUTE", "ROUTE", "ROUTE", "ROUTE", "ROUTE"]
direction = "Nord"
annees = ["2008"]
stations = []
sous_stations = []
nom_stations.each {|nom|
    station = Station.find_by_nom(nom)
    stations << station
    sous_stations << station.children.where(:direction => direction)
}
toute_valeurs = []
sous_stations = sous_stations.flatten
stations = sous_stations if sous_stations.any?
x = 0
stations.each {|station|
    annees.each {|annee|
        debut_periode = DateTime.parse("#{annee} - 01 - 01 00:00:00")
        fin_periode = DateTime.parse("#{annee} - 12 - 31 23:59:59")
        moyennes = MesureHoraire.where(:station_id => station, :horodatage =>
            debut_periode..fin_periode).group('extract(hour from
            horodatage)').order('extract(hour from horodatage)') .average(:valeur)
        valeurs = moyennes.map{|k,v| v.to_f}
        toute_valeurs << valeurs
        g.data("Route #{x}", valeurs) if valeurs.any?
        x += 1
    }
}
```

Code D.1 (Suite)

```
28 toute_valeurs.flatten!.delete_if{|d| d.nil?}
29 max = (toute_valeurs.max/1000.0).ceil*1000.to_i+1000
30 g.maximum_value = max
31 g.minimum_value = 0
32 colonnes = {}
33 time = (0..23).step
34 time.each{|t| colonnes[t] = t.to_s}
35 g.labels = colonnes
36 g.write('ponts_2008.png')
```

ANNEXE E

Tests de performance

Code E.1: Script d'insertion d'un ensemble de données aléatoires

```
1 #!/usr/bin/env ./script/runner
2
3 helper = ApplicationController.helpers
4 #Un lot doit être créé afin de répondre aux contraintes de la table mesures
5 lot = Lot.new(:usager => 'Test de performance!', :horodatage => DateTime.now(),
      :fichiers_source => 'Ensemble aléatoire')
6 lot.save
7
8 if Mesure.any?
9     #Seulement stations, paramètres, types et unités qui sont déjà utilisés par des
          mesures sont acquis afin de créer un environnement constant ou les
          contraintes devront être vérifiées
10    stations = Mesure.stations.uniq
11    parametres = Mesure.parametres.uniq
12    types = Mesure.types.uniq
13    unites = Mesure.unites
14 else
15    #Si aucune mesure n'est présente, une liste complète est acquise (ce script ne
          devrait être exécutés uniquement si des stations, paramètres, types et
          unités existent déjà)
16    stations = Station.all
17    parametres = Parametre.all
18    types = Type.all
19    unites = Unite.all
20 end
21 #Le temps de départ ainsi qu'une variable d'horodatage à insérer dans la base de
      données sont générés
22 temps_depart = Time.now()
23 temps_actuel = DateTime.now()
```

Code E.1 (Suite)

```
24 #Une boucle tournera pour 25 000 itérations afin de créer 25 000 nouvelles entrées
       aléatoires
25 25000.times do
26     #Un objet parent à chaque mesure doit être choisi, au hasard...
27     station = stations.sample
28     parametre = parametres.sample
29     type = types.sample
30     unite = unites.sample
31     #1 seconde est ajoutée au temps de départ afin de ne pas courir de risque de ne
           pas répondre aux contraintes de mesures
32     temps_actuel += 1.second
33     #Une valeur à insérer est générée de façon aléatoire
34     valeur = rand(100)
35     #Un nouvel objet mesure est créé et sauvegardé
36     m = Mesure.new(:station => station, :parametre => parametre, :type => type,
           :unite => unite, :lot => lot, :horodatage => temps_actuel, :valeur => valeur)
37     m.save
38 end
39 #Le temps de fin d'exécution est obtenu, et imprimé à l'écran
40 temps_fin = Time.now()
41 puts "Le␣temps␣d'exécution␣pour␣25000␣entrées␣aléatoires␣est␣de␣
       #{temps_fin-temps_depart}␣secondes."
```